Méthodes numériques géométriques et multi-échelles pour les équations différentielles (in English)

Gilles Vilmart

HDR / ENS CACHAN - BRETAGNE
sous le sceau de l'Université européenne de Bretagne

pour obtenir
L'HABILITATION À DIRIGER DES RECHERCHES
Mention : Mathématiques

présentée par
Gilles Vilmart
Préparée à l'Unité Mixte de Recherche 6625
Institut de recherche matématiques de Rennes

Méthodes numériques géométriques et multi-échelles pour les équations différentielles

Habilitation soutenue le 2 juillet 2013

Composition du jury :

François ALOUGES
Professeur - CMAP, École Polytechnique (Palaiseau) / *rapporteur*
Arieh ISERLES
Professeur - DAMTP, University of Cambridge (UK) / *rapporteur*
Claude LE BRIS
Ingénieur en chef des Ponts, des Eaux et des Forêts - MICMAC, CERMICS, ENPC (Marne-la-Vallée) / *rapporteur*

Philippe CHARTIER
Directeur de Recherche - IPSO, INRIA, ENS Cachan Bretagne et IRMAR (Rennes) / *examinateur*
Arnaud DEBUSSCHE
Professeur - ENS Cachan Bretagne, IPSO, INRIA et IRMAR (Bruz) / *examinateur*
Laurence HALPERN
Professeur - Université Paris 13, LAGA (Villetaneuse) / *examinatrice*
Patrick JOLY
Directeur de Recherche - POems, INRIA (Saclay) / *examinateur*

THÈSE

D'HABILITATION À DIRIGER DES RECHERCHES

présentée à

L'École Normale Supérieure de Cachan

Spécialité : mathématiques

Méthodes numériques géométriques et multi-échelles pour les équations différentielles

par

Gilles Vilmart

Version du 25 juin 2013

Remerciements

Je tiens tout d'abord à remercier François Alouges, Arieh Iserles et Claude Le Bris pour avoir accepté la charge de rapporteur, pour l'intérêt qu'ils ont bien voulu porter à mes travaux et pour leur participation au jury.

J'exprime ma sincère gratitude à Philippe Chartier, Arnaud Debussche, Laurence Halpern et Patrick Joly qui me font l'honneur de participer au jury, et pour les discussions que nous avons pu avoir.

Je désire remercier chaleureusement mes directeurs de thèse, Philippe Chartier et Ernst Hairer, pour leur amitié, l'intérêt porté à mes travaux, leur soutien et leur disponibilité.

Je souhaite remercier vivement toutes les personnes avec lesquelles j'ai eu l'occasion de travailler et en particulier mes co-auteurs : Assyr Abdulle, Yun Bai, François Castella, Philippe Chartier, Monique Chyba, David Cohen, Stéphane Descombes, Ernst Hairer, Joseba Makazaga, Ander Murua, Konstantinos Zygalakis. J'espère que nous aurons encore de nombreuses occasions de travailler ensemble dans le futur.

Enfin, un grand merci aux membres de l'équipe IPSO de l'INRIA, du département de mathématiques de l'ENS Rennes et de l'IRMAR car ils ont rendu très agréables ces deux dernières années passées à Rennes.

à Shaula, à Matteo

à Thierry

Contents

Introduction

This manuscript constitutes a synthesis document of my research in preparation for my habilitation degree in Mathematics.

Since September 2011, I am an assistant professor (a long-term nine year position called "agrégé-préparateur") in the department of Mathematics of École Normale Supérieure (ENS) de Cachan, Brittany extension. I am member of the IPSO team of INRIA Rennes headed by Philippe Chartier, which belongs to the numerical analysis group of the Institut de recherche mathématique de Rennes (IRMAR) headed by Florian Méhats.

I obtained my Ph.D. in 2008 in both the University of Geneva and the University of Rennes 1, in the frame of a double doctorate program ("cotutelle internationale"), under the joint direction of Ernst Hairer (Geneva) and Philippe Chartier (IPSO, Rennes). Right after my Ph.D. on geometric numerical integration of differential equations, I decided to discover new research topics on multiscale problems. I initiated a fruitful collaboration as a post-doc with Assyr Abdulle and his group (with Yun Bai and Martin Huber) at École Polytechnique Fédérale de Lausanne, Chair of Computational and Applied Mathematics, first on multiscale finite element methods for homogenization in PDEs, and second on geometric and multiscale time integrators for stochastic problems (together also with David Cohen, Umeå University, and Konstantinos Zygalakis, Southampton University). In the last years, a growing interest of the IPSO team in multiscale and oscillatory PDEs and S(P)DEs has also independently arisen, and it was naturally that I joint back the IPSO team and initiated new collaborations (also with Ander Murua and Joseba Makazaga, San Sebastian).

My research focuses on the numerical analysis of geometric and multiscale integrators for deterministic or stochastic differential equations. Numerous physical (or chemical) phenomena can be modeled by differential equations which possess a particular geometric or multiscale structure (e.g. Hamiltonian structures, Poisson structures, first integrals, multiscale structure in time or in space, highly oscillatory systems), but their complexity is often so huge that a satisfactory solution is out of reach using only general purpose numerical integrators, e.g. a high-order explicit Runge-Kutta method in time or a standard finite element method with a very fine mesh in space. The aim is thus to identify the relevant geometric or multiscale properties of such problems, and try to take advantage of them to design and study new efficient and reliable integrators that reproduce the qualitative behaviour of the exact solution of the considered models.

This documents is organized into three chapters, each of them corresponding to a research orientation in the geometric and multiscale integration of differential equations. Each chapter starts with a short review of background material to make the manuscript reasonably self-contained and the strong relations and unity between them is stressed all along. Perspectives including ongoing and future work conclude each chapter.

The first chapter is devoted to the theme of geometric numerical integration of differential equations which was at the core of my Ph.D. work and has links with a priori distant

fields of Mathematics (Combinatorial algebra and renormalization in Quantum field theory). Inspired by recent advances in the theory of modified differential equations, a main result of my Ph.D. was a general framework for the construction of high-order geometric integrators for deterministic ODEs, and a spectacular application of this theory is the design of the first high-order geometric integrator for the free rigid body motion (and which is not derived using a standard composition technique). This approach is generalized to stochastic differential equations (SDEs) and is illustrated with the constructions of new methods of weak order two, in particular, implicit integrators that exactly conserve all quadratic first integrals of a stochastic dynamical system, and also (semi-)implicit integrators well suited for stiff (mean-square stable) stochastic problems, which is the target of the second chapter. A recent work based on the geometric idea of composition methods concludes this first chapter. Although composition methods are originally designed for non-stiff problems, we introduce a new class multi-revolution composition methods for highly oscillatory problems in time, with applications including the long time integration of the nonlinear Schrödinger equation. These integrators can be cast in the multiscale framework of micro-macro integrator (here in time), in the spirit of the homogenization methods for multiscale (in space) PDEs discussed in the third chapter. This also illustrate well the complementarity of the research topics addressed in this thesis.

The second chapter focuses on the construction of weak high-order integrators for stiff (mean-square stable) SDEs in general dimensions with a general non-commutative noise. Similarly to the deterministic case, it is in general difficult to construct integrators that are both accurate and with favorable stability properties to avoid severe time step restrictions in the integration of stiff SDEs, due to the variety of the scales involved. By stabilizing an efficient derivative-free version of the standard Milstein-Talay method of weak second order, we introduce a Runge-Kutta type integrator that has simultaneously weak second order of accuracy and the mean-square A-stability property. This integrator is semi-implicit: it is implicit with respect to the deterministic part, but explicit with respect to the stochastic part. We next introduce a family of explicit stabilized integrators with weak order two with extended mean-square stability domains, which are of great interest for large systems SDEs (arising e.g. from an SPDEs spatial discretization). These integrators of weak order two are to the best of our knowledge the first of their kind with such favorable mean-square stability properties. These ideas inspired the construction of a new "swiss-knife" integrator that permits to integrate stiff diffusion-advection-reaction problems with (or without) stochastic noise in various regimes of stiffness and Peclet numbers with all the advantages of explicit stabilized integrators.

The third chapter concerns the construction and analysis of multiscale integrators for homogenization in PDEs. We focus on the so-called finite element Heterogeneous Multiscale method (FE-HMM) which relies on the approach of multi-grid methods by coupling finite element methods at the microscopic and macroscopic scales. This method permits to drastically reduce the number of degrees of freedom compared to standard finite element methods. We present an a priori error analysis with optimal convergence rates in the H^1 and L^2 norms for two classes of problems with quite a different nature but a certain unified framework of their analysis: first multiscale linear parabolic problems with a time-dependent tensor, and second quasilinear elliptic problems of nonmonotone type. The coupling of the FE-HMM for parabolic problems with time-integrators with favourable stability properties is addressed (implicit and explicit stabilized integrators, as studied in Chapter 2 in the context of stochastic problems), with an analysis of the fully discrete time–micro–macro discretizations, while the analysis of the nonlinear FE-HMM for nonmonotone elliptic problems is the first analysis proposed in the literature that is valid in dimension three of space.

Introduction (French)

Ce document constitue une synthèse de mes travaux de recherche en vue d'obtenir l'habilitation à diriger des recherches en mathématiques.

Je suis, depuis septembre 2011, agrégé-préparateur dans le département de mathématiques de l'École Normale Supérieure (ENS) de Cachan, antenne de Bretagne. Je suis membre de l'équipe IPSO de l'INRIA Rennes, dirigée par Philippe Chartier, et appartenant à l'équipe d'analyse numérique de l'Institut de recherche en mathématiques de Rennes (IRMAR), dirigée par Florian Méhats.

J'ai obtenu mon doctorat de mathématiques en 2008 dans le cadre d'une thèse en cotutelle internationale entre l'Université de Genève et l'Université de Rennes 1, sous la direction conjointe de Ernst Hairer (Genève) et Philippe Chartier (IPSO, Rennes). Après ma thèse sur l'intégration numérique géométrique des équations différentielles ordinaires (EDO) et partielles (EDP), j'ai décidé de découvrir de nouvelles thématiques sur les problèmes multi-échelles en temps et espace.

J'ai initié une collaboration fructueuse en tant que post-doc avec Assyr Abdulle et son groupe (notamment Yun Bai et Martin Huber) à l'École Polytechnique Fédérale de Lausanne, Chaire d'analyse numérique, d'une part sur les méthodes multi-échelles de type éléments finis pour l'homogénéisation dans les EDP, et d'autre part sur les intégrateurs géométriques et multi-échelles en temps pour les équations différentielles stochastiques (EDS) (avec également David Cohen, Université de Umeå, et Konstantinos Zygalakis, Université de Southampton). Ces dernières années un intérêt croissant de l'équipe IPSO pour les problèmes multi-échelles d'EDO et d'EDP hautement oscillantes (et stochastiques) a également émergé, et c'est naturellement que j'ai réintégré l'équipe IPSO et initié de nouvelles collaborations (également avec Ander Murua et Joseba Makazaga, San Sebastian).

Ma recherche porte sur l'analyse numérique des intégrateurs géométriques et multi-échelles pour les équations différentielles déterministes ou stochastiques. De nombreux phénomènes physiques (ou chimiques) peuvent être modélisés par des équations différentielles qui possèdent une structure géométrique ou multi-échelles particulière (par exemple, les structures hamiltoniennes, les structures de Poisson, les intégrales premières, des structures multi-échelles en temps ou en espace, des systèmes hautement oscillatoires), mais leur complexité est souvent telle qu'une solution satisfaisante est hors de portée en utilisant seulement des intégrateurs numériques standards à usage général, par exemple, une méthode de Runge-Kutta explicite d'ordre élevé en temps ou une méthode d'éléments finis standard avec un maillage fin en espace. L'objectif est donc d'identifier les propriétés géométriques ou multi-échelles pertinentes de ces problèmes, et d'essayer d'en tirer avantage pour concevoir et étudier de nouveaux intégrateurs efficaces et fiables, qui reproduisent fidèlement le comportement qualitatif de la solution exacte des modèles considérés.

Cette thèse est organisée en trois chapitres, chacun correspondant à une thématique de recherche pour l'intégration géométrique et multi-échelles des équations différentielles. Chaque chapitre débute par un bref rappel des outils nécessaires et les liens entre ces

différents thèmes sont soulignés. De brèves perspectives sur les travaux en cours et à venir concluent chaque chapitre.

Le premier chapitre est consacré à l'intégration numérique géométrique des équations différentielles, le sujet au cœur de mon travail de doctorat, et qui possède des liens avec d'autres champs des mathématiques a priori éloignés (algèbre combinatoire et renormalisation en théorie quantique des champs). Inspiré par les progrès récents dans la théorie des équations différentielles modifiées, une contribution importante de mes travaux de doctorat est une méthodologie pour la construction d'intégrateurs géométriques d'ordre élevé de convergence pour les EDO déterministes, et une application exemplaire de cette théorie est la construction du premier intégrateur géométrique d'ordre élevé pour la dynamique d'un corps rigide libre (non obtenue par la technique standard des méthodes de compositions). Cette approche est généralisée aux équations différentielles stochastiques (EDS) et est illustrée par la construction de nouvelles méthodes d'ordre faible deux, en particulier, des intégrateurs implicites qui conservent exactement toutes les intégrales premières quadratiques d'un système stochastique, et aussi des intégrateurs (semi)-implicites bien adaptés aux problèmes stochastiques raides (en sens des moyennes quadratiques), thème objet du second chapitre. Un travail récent basé sur l'idée géométrique des méthodes de composition conclut ce chapitre. Bien que les méthodes de composition sont à l'origine conçues pour les problèmes non raides, nous introduisons une nouvelle classe de méthodes de composition multi-révolutions pour les problèmes hautement oscillatoires en temps, avec application notamment à l'équation de Schrödinger non-linéaire. Ces intégrateurs s'inscrivent dans le cadre des méthodes multi-échelles micro-macro (ici en temps), dans l'esprit des méthodes d'homogénéisation d'EDP multi-échelles (en espace) présenté dans le troisième chapitre. Ceci illustre aussi la complémentarité des thèmes de recherche abordés dans cette thèse.

Le second chapitre se focalise sur la construction de méthodes d'ordre élevé au sens faible pour les problèmes d'EDS raides (et stables en moyenne quadratique) avec un bruit général non commutatif. A l'instar du cas déterministe, il est en général difficile de construire des intégrateurs qui soient à la fois d'ordre élevé et avec des propriétés de stabilité favorables pour éviter des restrictions sévères de longueur de pas de temps pour les problèmes d'EDS raides. En stabilisant une variante efficace de la méthode de Milstein-Talay d'ordre faible deux, nous introduisons une méthode de type Runge-Kutta stochastique qui est à la fois d'ordre faible deux et A-stable au sens stochastique. Cet intégrateur est semi-implicite : il est implicite par rapport à la partie déterministe, mais explicite par rapport à la partie stochastique de l'EDS. Nous introduisons ensuite une famille d'intégrateurs explicites stabilisés d'ordre faible deux avec de grand domaines de stabilité (au sens des moyennes quadratiques), qui sont d'un grand intérêt pour les systèmes stochastiques raides en grande dimension (par exemple issus d'une discrétisation spatiale d'EDP stochastique). Ces intégrateurs d'ordre faible deux sont à notre connaissance les premiers construits avec d'aussi bonnes propriétés de stabilité. Ces idées ont inspiré la construction d'un nouvel intégrateur dit "couteau-suisse" qui permet d'intégrer des problèmes raides de type diffusion-advection-réaction avec (ou sans) bruit stochastique pour une grande variétés de régimes de raideur et de nombres de Péclet, et tout en conservant les avantages des intégrateurs explicites stabilisés.

Le troisième chapitre porte sur la construction et l'analyse d'intégrateurs multi-échelles pour l'homogénéisation dans les EDP. Nous nous concentrons sur la méthode des éléments finis hétérogène multi-échelles (FE-HMM) qui s'appuie sur le principe des méthodes multi-grilles en couplant des méthodes d'éléments finis aux échelles microscopique et macroscopique. Cette méthode permet de réduire considérablement le nombre de degrés de

liberté par rapport aux méthodes d'éléments finis standards. Nous présentons une analyse a priori de l'erreur avec des vitesses de convergences optimales dans les normes L^2 et H^1 pour deux classes de problèmes de natures différentes, mais avec un cadre unifié pour leur analyse multi-échelles : dans un premier temps des problèmes paraboliques linéaires avec un tenseur hautement oscillant en espace et dépendant du temps, et dans un second temps des problèmes elliptiques quasi-linéaires de type non monotone. Une originalité du travail est l'étude du couplage de la FE-HMM pour des problèmes paraboliques avec des intégrateurs en temps ayant de bonnes propriétés de stabilité (intégrateurs implicites et explicites stabilisés, comme étudié dans un contexte stochastique dans le second chapitre), avec une analyse de la discrétisation complètes temps/échelle micro/échelle macro. L'analyse non linéaire pour des problèmes elliptiques non monotones est la première analyse de la FE-HMM proposée qui soit valable en dimension trois d'espace.

Chapter 1

High-order geometric methods for deterministic and stochastic problems

We first recall in Section 1.1 important classical tools in the context of geometric numerical integration, together with some Ph.D. contributions of the author, and that will be useful to understand the results presented, in particular an extension to stochastic differential equations (Section 1.2), and a new class of geometric integrators for highly-oscillatory problems obtained by considering modified oscillatory periods (Section 1.3).

1.1 Modified differential equations

Modified differential equations in combination with backward error analysis form an important tool for studying the long-time behaviour of numerical integrators for ordinary differential equations (cf. the monographs [HLW06] and [LR04]). The main idea of this theory is sketched and, by inverting the roles of the exact and numerical flows, a new approach for the construction of high order numerical integrators for ordinary differential equations is developed [CHV07b, CHV07a, Vil08a]. As an application, a computationally efficient and highly accurate modification of the Discrete Moser–Veselov algorithm for the simulation of the free rigid body is presented [HV06].

1.1.1 Modified equations for backward error analysis

Consider an initial value problem

$$\dot{y} = f(y), \qquad y(0) = y_0 \tag{1.1}$$

with sufficiently smooth vector field $f(y)$, and a numerical one-step integrator $y_{n+1} = \Phi_{f,h}(y_n)$. The idea of backward error analysis is to search for a modified differential equation

$$\dot{z} = f_h(z) = f(z) + hf_2(z) + h^2 f_3(z) + \ldots, \qquad z(0) = y_0, \tag{1.2}$$

which is a formal series in powers of the step size h, such that the numerical solution $\{y_n\}$ is formally equal to the exact solution of (1.2),

$$y_n = z(nh) \qquad \text{for} \quad n = 0, 1, 2, \ldots, \tag{1.3}$$

see the top picture of Figure 1.1.

The idea of backward error analysis was originally introduced by Wilkinson (1960) in the context of numerical linear algebra. For the integration of ODEs it was not used until one became interested in the long-time behaviour of numerical solutions. Without considering it as a theory, Ruth [Rut83] uses the idea of backward error analysis to motivate symplectic integrators for Hamiltonian systems. In fact, applying a symplectic numerical method to a Hamiltonian system

$$\dot{q} = \nabla_p H(p, q), \quad \dot{p} = -\nabla_q H(p, q),$$

gives rise to a modified differential equation that is also Hamiltonian. This is the main tool for proving the good conservation (without drift) of the energy by symplectic integrators applied to Hamiltonian systems over (exponentially) long time intervals (under appropriate assumptions). Indeed, it permits to transfer known properties of perturbed Hamiltonian systems (e.g., conservation of energy, KAM theory for integrable systems) to properties of symplectic numerical integrators. One became soon aware that this kind of reasoning is not restricted to Hamiltonian systems, and new insight can be obtained with the same techniques also for reversible differential equations, for Poisson systems, for divergence-free problems, etc. A rigourous analysis has been developed in the nineties. We refer

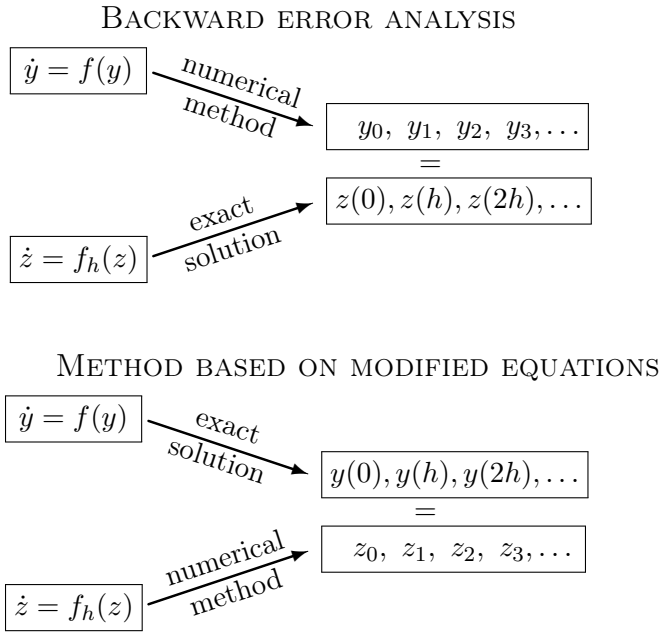


Figure 1.1: Backward error analysis opposed to numerical integrators based on modified equations

the interested reader to [HLW06, Chapter IX], where backward error analysis and its applications are explained in detail (see also the study [CHV09] in the context of optimal control).

1.1.2 Numerical integrators based on modified equations

Backward error analysis is a purely theoretical tool that gives much insight into the long-term integration with geometric numerical methods. We shall show that by simply exchanging the roles of the "numerical method" and the "exact solution" (cf. the two pictures in Figure 1.1), it can be turned into a mean for constructing high order integrators that conserve geometric properties. They will be useful for integrations over long times.

Let us be more precise. As before, we consider an initial value problem (1.1) and a numerical integrator. But now we search for a modified differential equation, again of the form (1.2), such that the numerical solution $\{z_n\}$ of the method applied with step size h to (1.2) yields formally the exact solution of the original differential equation (1.1), i.e.,

$$z_n = y(nh) \qquad \text{for} \quad n = 0, 1, 2, \ldots, \tag{1.4}$$

see the bottom picture of Figure 1.1. Note that this modified equation is different from the one considered before. However, due to the close connection with backward error analysis, all theoretical and practical results have their analogue in this new context. The modified differential equation is again an asymptotic series that usually diverges, and its truncation inherits geometric properties of the exact flow if a suitable integrator is applied. The coefficient functions $f_j(z)$ can be computed recursively by using a formula manipulation program like MAPLE. This can be done by developing both sides of $z(t+h) = \Phi_{f_h,h}(z(t))$ into a series in powers of h, and by comparing their coefficients. Once a few functions $f_j(z)$ are known, the following algorithm suggests itself.

Algorithm 1.1.1 (Numerical integrators based on modified equations) *Consider the truncation*

$$\dot{z} = f_h^{[r]}(z) = f(z) + hf_2(z) + \cdots + h^{r-1}f_r(z) \tag{1.5}$$

of the modified differential equation corresponding to $\Phi_{f,h}(y)$*. Then,*

$$z_{n+1} = \Psi_{f,h}(z_n) := \Phi_{f_h^{[r]},h}(z_n)$$

defines a numerical method of order r *that approximates the solution of (1.1). We call it integrator based on modified equations, because the vector field* $f(y)$ *of (1.1) is modified into* $f_h^{[r]}$ *before the basic integrator is applied.*

This is an alternative approach for constructing high order numerical integrators for ODEs (classical approaches are multistep, Runge–Kutta, Taylor series, extrapolation, composition, and splitting methods). It is particularly interesting in the context of geometric integration because, as known from backward error analysis, the modified differential equation inherits the same structural properties as (1.1) if a suitable integrator is applied.

A few known methods can be cast into the framework of integrators based on modified equations although they have not been constructed in this way. The most important are the generating function methods as introduced by Feng [Fen86]. These are high order symplectic integrators obtained by applying a simple symplectic method to a modified Hamiltonian system. The corresponding Hamiltonian is the solution of a Hamilton–Jacobi partial differential equation. Another special case is a modification of the discrete Moser–Veselov algorithm for the Euler equations of the rigid body, proposed by McLachlan and Zanna [MZ05]. The general approach of Algorithm 1.1.1 and the example of Algorithm 1.1.2 are introduced and discussed in [CHV07b].

Algorithm 1.1.2 *For the numerical integration of (1.1) we consider the implicit midpoint rule*

$$y_{n+1} = y_n + h\, f\Big(\frac{y_n + y_{n+1}}{2}\Big). \tag{1.6}$$

Applying (1.6) *to* $\dot{z} = f_h^{[5]}(z)$ *with the truncated modified vector given by*

$$\begin{aligned} f_h^{[5]} \;=\; & f \;+\; \frac{h^2}{12}\Big(-f'f'f + \frac{1}{2}f''(f,f)\Big) \;+\; \frac{h^4}{120}\Big(f'f'f'f'f - f''(f,f'f'f) + \frac{1}{2}f''(f'f,f'f)\Big) \\ & +\; \frac{h^4}{240}\Big(-\frac{1}{2}f'f'f''(f,f) + f'f''(f,f'f) + \frac{1}{2}f''(f,f''(f,f)) - \frac{1}{2}f^{(3)}(f,f,f'f)\Big) \\ & +\; \frac{h^4}{80}\Big(-\frac{1}{6}f'f^{(3)}(f,f,f) + \frac{1}{24}f^{(4)}(f,f,f,f)\Big) \end{aligned} \tag{1.7}$$

yields a numerical approximation of order 6 *for (1.1) which is symmetric (i.e.* $\Phi_h \circ \Phi_{-h}(y) = y$*). In addition, it is symplectic for all Hamiltonian vector field* f*.*

Here, we use the notations $f'(x)\cdot$ for the first derivative (a linear form) and $f''(x)(\cdot,\cdot)$ for the second derivative (a symmetric bilinear form) of f at the point X_0 and similar notations for the higher order derivatives. At first glance the modified equation (1.7) looks extremely complicated and it is hard to imagine that the modified midpoint rule can compete with other methods of the same order. This is true in general, but there are important differential equations for which the evaluation of $f_h^{[r]}(y)$ is not much more expensive than that of $f(y)$, so that the modified integrators of Algorithm 1.1.1 can become efficient. A spectacular example is the equations of motion for the full dynamics of a rigid body (see [CHV07b, HV06, Vil08b], the survey [Vil13], and Section 1.1.2 below).

Accurate rigid-body integrator based on the DMV algorithm As illustration of how efficient integrators based on modified equations can be, we consider the equations of motion for a rigid body,

$$\dot{y} = \widehat{y}\,\mathcal{I}^{-1}y, \qquad \dot{Q} = Q\,\widehat{\mathcal{I}^{-1}y}, \qquad \text{where} \qquad \widehat{a} = \begin{pmatrix} 0 & -a_3 & a_2 \\ a_3 & 0 & -a_1 \\ -a_2 & a_1 & 0 \end{pmatrix} \tag{1.8}$$

for a vector $a = (a_1, a_2, a_3)^T$. Here, $\mathcal{I} = \mathrm{diag}(I_1, I_2, I_3)$ is the matrix formed by the moments of inertia, y is the vector of the angular momenta, and Q is the orthogonal matrix that describes the rotation relative to a fixed coordinate system. As numerical integrator we choose the Discrete Moser–Veselov algorithm (DMV) [MV91],

$$\widehat{y}_{n+1} = \Omega_n\,\widehat{y}_n\,\Omega_n^T, \qquad Q_{n+1} = Q_n\,\Omega_n^T, \tag{1.9}$$

where the orthogonal matrix Ω_n is given from $\Omega_n^T D - D\,\Omega_n = h\,\widehat{y}_n$. Here, the diagonal matrix $D = \mathrm{diag}(d_1, d_2, d_3)$ is determined by $d_1 + d_2 = I_3$, $d_2 + d_3 = I_1$, and $d_3 + d_1 = I_2$. This algorithm is an excellent geometric integrator and shares many geometric properties with the exact flow. It is symplectic, it exactly preserves the Hamiltonian, the Casimir and the angular momentum Qy (in the fixed frame), and it keeps the orthogonality of Q, which permits an efficient implementation using quaternions. Its only disadvantage is the low order two.

The technique of integrators based on modified equations cannot be directly applied to increase the order of this method, because the algorithm (1.9) is not defined for general problems (1.1). It is, however, defined for arbitrary I_j, and therefore we look for modified moments of inertia $\widetilde{I}_j$ such that the DMV algorithm applied with $\widetilde{I}_j$ yields the exact solution of (1.8). It is shown in [HV06] that this is possible with

$$\frac{1}{\widetilde{I}_j} = \frac{1}{I_j}\Big(1 + h^2 s_3(y_n) + h^4 s_5(y_n) + \cdots\Big) + h^2 d_3(y_n) + h^4 d_5(y_n) + \cdots\,. \tag{1.10}$$

The expressions $s_k(y)$ and $d_k(y)$ can be computed by a formula manipulation package similar as the modified differential equation is obtained. The first of them are

$$\begin{aligned} s_3(y_n) &= -\frac{1}{3}\left(\frac{1}{I_1} + \frac{1}{I_2} + \frac{1}{I_3}\right) H(y_n) + \frac{I_1 + I_2 + I_3}{6\,I_1\,I_2\,I_3}\,C(y_n), \\ d_3(y_n) &= \frac{I_1 + I_2 + I_3}{6\,I_1\,I_2\,I_3}\,H(y_n) - \frac{1}{3\,I_1\,I_2\,I_3}\,C(y_n), \end{aligned}$$

where

$$C(y) = \frac{1}{2}\Big(y_1^2 + y_2^2 + y_3^2\Big) \qquad \text{and} \qquad H(y) = \frac{1}{2}\Big(\frac{y_1^2}{I_1} + \frac{y_2^2}{I_2} + \frac{y_3^2}{I_3}\Big) \tag{1.11}$$

are the Casimir and the Hamiltonian of the system. The physical interpretation of this result is the following: after perturbing suitably the form of the body, an application of the DMV algorithm yields the exact motion of the body. Truncating the series in (1.10) after the h^{2r-2} terms, yields a modified DMV algorithm of order $2r$.

1.1.3 Algebraic structures of Butcher-series

Since the work of Cayley [Cay57] and Merson [Mer57] it is known that the expressions arising in the derivatives of the solution of (1.1), $\ddot{y} = (f'f)(y)$, $\dddot{y} = (f''(f,f))(y) +$

$(f'f'f)(y)$, are in one-to-one correspondence with rooted trees. It is therefore natural to consider formal series of the form

$$\begin{aligned} B(a,hf,y) &= a(\emptyset)y + ha(\bullet)f(y) + h^2a(\bullet\!\!-\!\!\bullet)(f'f)(y) + \frac{h^3}{2}a(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet)\big(f''(f,f)\big)(y) \\ &+ h^3a(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet)(f'f'f)(y) + h^4a(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet)\big(f''(f,f'f)\big)(y) + \dots \end{aligned} \tag{1.12}$$

with scalar coefficients $a(\emptyset)$, $a(\bullet)$, $a(\bullet\!\!-\!\!\bullet)$, etc. The exact solution of (1.1) is of this form with $a(\emptyset) = a(\bullet) = 1$, $a(\bullet\!\!-\!\!\bullet) = 1/2$, $a(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) = 1/3$, etc. In his fundamental work on order conditions, Butcher discovered in the 1960ies (culminating in the seminal article [But72]) that the numerical solution of a Runge–Kutta method is also a series of the form (1.12) with $a(\tau)$ depending only on the coefficients of the method. Hairer and Wanner [HW74] considered series (1.12) with arbitrary coefficients and called them B-series[1]. They applied them to the elaboration of order conditions for general multi-value methods. B-series and extensions thereof are now exposed in various textbooks and articles, possibly with different normalizations e.g., [HLW06, But08].

B-series play an important role in the study and construction of numerical integrators. This is a consequence of the following two operations on B-series:

- *Composition law* ([But72, HW74]). For $b(\emptyset) = 1$, a B-series considered as a mapping $y \mapsto B(b,hf,y)$ is $\mathcal{O}(h)$-close to the identity. It is therefore possible to replace y in (1.12) with $B(b,hf,y)$, and to expand all expressions around y. Interestingly, the result is again a B-series and we have

$$B\big(a,hf,B(b,hf,y)\big) = B(b\cdot a,hf,y). \tag{1.13}$$

- *Substitution law* ([CHV05, CHV07b]). For $b(\emptyset) = 0$, the B-series $B(b,hf,y)$ is a vector field that is a perturbation of $hf(y)$, multiplied by the scalar $b(\bullet)$. Therefore, we can substitute the vector field $B(b,hf,\cdot)$ for hf in (1.12). Also in this case we obtain a B-series, which we denote

$$B\big(a,B(b,hf,\cdot),y\big) = B(b\star a,hf,y). \tag{1.14}$$

A straightforward computation yields for the composition law $(b\cdot a)(\emptyset) = a(\emptyset)$ and

$$\begin{aligned} (b\cdot a)(\bullet) &= a(\emptyset)b(\bullet) + a(\bullet), \\ (b\cdot a)(\bullet\!\!-\!\!\bullet) &= a(\emptyset)b(\bullet\!\!-\!\!\bullet) + a(\bullet)b(\bullet) + a(\bullet\!\!-\!\!\bullet), \\ (b\cdot a)(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) &= a(\emptyset)b(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) + a(\bullet)b(\bullet)^2 + 2a(\bullet\!\!-\!\!\bullet)b(\bullet) + a(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet), \\ (b\cdot a)(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) &= a(\emptyset)b(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) + a(\bullet)b(\bullet\!\!-\!\!\bullet) + a(\bullet\!\!-\!\!\bullet)b(\bullet) + a(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet). \end{aligned} \tag{1.15}$$

Similarly, for the substitution law we obtain $(b\star a)(\emptyset) = a(\emptyset)$ and

$$\begin{aligned} (b\star a)(\bullet) &= a(\bullet)b(\bullet), \\ (b\star a)(\bullet\!\!-\!\!\bullet) &= a(\bullet)b(\bullet\!\!-\!\!\bullet) + a(\bullet\!\!-\!\!\bullet)b(\bullet)^2, \\ (b\star a)(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) &= a(\bullet)b(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) + 2a(\bullet\!\!-\!\!\bullet)b(\bullet)b(\bullet\!\!-\!\!\bullet) + a(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet)b(\bullet)^3, \\ (b\star a)(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) &= a(\bullet)b(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet) + 2a(\bullet\!\!-\!\!\bullet)b(\bullet)b(\bullet\!\!-\!\!\bullet) + a(\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet)b(\bullet)^3. \end{aligned} \tag{1.16}$$

General formulae for the substitution law where first derived in [CHV05] (see also [CHV10]).

[1] Originally named Butcher series.

The composition law is an important tool for the construction of various integration methods, such as Runge–Kutta methods, general linear methods, Rosenbrock methods, multi-derivative methods, etc. It allows the derivation of the order conditions for arbitrarily high orders in an elegant way avoiding tedious series expansions [HNW93, HW96]. Another application is the composition of different numerical integrators yielding higher accuracy: effective order or pre- and post-processing of composition methods [But69, BCR99].

Applications of the substitution law are more recent and mainly in connection with structure-preserving algorithms (geometric numerical integration). This law gives much insight into the modified differential equation of backward error analysis [HLW06], and it is the main ingredient for the construction of integrators based on modified equations [CHV07b].

Group and monoid structures. Let $T = \{\bullet, \begin{picture}(8,8)\put(1,1){\line(0,1){6}}\put(1,7){\line(1,0){6}}\put(1,1){\circle*{1.5}}\put(7,7){\circle*{1.5}}\end{picture}, \begin{picture}(8,8)\put(1,1){\line(1,1){3}}\put(5,1){\line(-1,1){3}}\put(1,1){\circle*{1.5}}\put(5,1){\circle*{1.5}}\put(4,4){\circle*{1.5}}\end{picture}, \ldots\}$ be the set of rooted trees, and consider the set $T_0 = T \cup \{\emptyset\}$ including the empty tree. The set of mappings

$$G_C = \{a\colon T_0 \to \mathbb{R}\,;\, a(\emptyset) = 1\} \tag{1.17}$$

with the product (1.15) of the composition law is a group. Identity is the element that corresponds to the B-series $B(a, hf, y) = y$. Associativity follows from that of the composition of mappings and the existence of an inverse is obtained from the explicit formulae for the product. The group G_C has been introduced in [But72] and is called the Butcher group in [HW74].

In a similar way, the substitution law (1.16) makes the set

$$G_S = \{a\colon T_0 \to \mathbb{R}\,;\, a(\emptyset) = 0\} \tag{1.18}$$

a monoid. It is a monoid of vector fields and has first been considered in [CHV05]. The identity element is the mapping that corresponds to the B-series $B(a, hf, y) = hf(y)$. Invertible elements in G_S are those with $a(\bullet) \neq 0$ and yield the group

$$G_S^* = \{a\colon T_0 \to \mathbb{R}\,;\, a(\emptyset) = 0,\, a(\bullet) \neq 0\}. \tag{1.19}$$

Hopf algebras of trees. Independently of the theory of B-series, Connes and Moscovici [CM98] in the context of non-commutative geometry, and Connes and Kreimer [CK98, CK00] in the theory of renormalization consider a Hopf algebra of rooted trees whose co-product is for the first trees given by $\Delta_{CK}(\emptyset) = \emptyset \otimes \emptyset$ and

$$\begin{aligned}
\Delta_{CK}(\bullet) &= \bullet \otimes \emptyset + \emptyset \otimes \bullet, \\
\Delta_{CK}(\begin{picture}(8,8)\put(1,1){\line(0,1){6}}\put(1,7){\line(1,0){6}}\put(1,1){\circle*{1.5}}\put(7,7){\circle*{1.5}}\end{picture}) &= \begin{picture}(8,8)\put(1,1){\line(0,1){6}}\put(1,7){\line(1,0){6}}\put(1,1){\circle*{1.5}}\put(7,7){\circle*{1.5}}\end{picture} \otimes \emptyset + \bullet \otimes \bullet + \emptyset \otimes \begin{picture}(8,8)\put(1,1){\line(0,1){6}}\put(1,7){\line(1,0){6}}\put(1,1){\circle*{1.5}}\put(7,7){\circle*{1.5}}\end{picture}, \\
\Delta_{CK}(\begin{picture}(8,8)\put(1,1){\line(1,1){3}}\put(5,1){\line(-1,1){3}}\put(1,1){\circle*{1.5}}\put(5,1){\circle*{1.5}}\put(4,4){\circle*{1.5}}\end{picture}) &= \begin{picture}(8,8)\put(1,1){\line(1,1){3}}\put(5,1){\line(-1,1){3}}\put(1,1){\circle*{1.5}}\put(5,1){\circle*{1.5}}\put(4,4){\circle*{1.5}}\end{picture} \otimes \emptyset + \bullet\bullet \otimes \bullet + 2\,\bullet \otimes \begin{picture}(8,8)\put(1,1){\line(0,1){6}}\put(1,7){\line(1,0){6}}\put(1,1){\circle*{1.5}}\put(7,7){\circle*{1.5}}\end{picture} + \emptyset \otimes \begin{picture}(8,8)\put(1,1){\line(1,1){3}}\put(5,1){\line(-1,1){3}}\put(1,1){\circle*{1.5}}\put(5,1){\circle*{1.5}}\put(4,4){\circle*{1.5}}\end{picture}, \\
\Delta_{CK}(\begin{picture}(8,8)\put(1,1){\line(0,1){6}}\put(1,7){\line(1,0){6}}\put(7,7){\line(0,-1){6}}\put(1,1){\circle*{1.5}}\put(7,7){\circle*{1.5}}\put(7,1){\circle*{1.5}}\end{picture}) &= \begin{picture}(8,8)\put(1,1){\line(0,1){6}}\put(1,7){\line(1,0){6}}\put(7,7){\line(0,-1){6}}\put(1,1){\circle*{1.5}}\put(7,7){\circle*{1.5}}\put(7,1){\circle*{1.5}}\end{picture} \otimes \emptyset + \begin{picture}(8,8)\put(1,1){\line(0,1){6}}\put(1,7){\line(1,0){6}}\put(1,1){\circle*{1.5}}\put(7,7){\circle*{1.5}}\end{picture} \otimes \bullet + \bullet \otimes \begin{picture}(8,8)\put(1,1){\line(0,1){6}}\put(1,7){\line(1,0){6}}\put(1,1){\circle*{1.5}}\put(7,7){\circle*{1.5}}\end{picture} + \emptyset \otimes \begin{picture}(8,8)\put(1,1){\line(0,1){6}}\put(1,7){\line(1,0){6}}\put(7,7){\line(0,-1){6}}\put(1,1){\circle*{1.5}}\put(7,7){\circle*{1.5}}\put(7,1){\circle*{1.5}}\end{picture}.
\end{aligned} \tag{1.20}$$

Brouder [Bro00, Bro04] (and also implicitly Dür [Dür86]) noticed the close connection between this co-product and the product (1.15) of the composition law.

Indeed, it is obtained from (1.15) by writing the argument of the mapping a to the right of the $\otimes$ sign, and those of the mapping b to the left of it. To the last terms in (1.15), which do not contain any $b(\tau)$, one adds the trivial factor $b(\emptyset) = 1$.

It is not surprising that a similar connection holds also for the substitution law. Inspired by the work [CHV05], Calaque, Ebrahimi-Fard and Manchon [CEFM09] introduced a coproduct which, for the first trees, is given by

$$\begin{aligned}
\Delta_{CEM}(\bullet) &= \bullet\otimes\bullet,\\
\Delta_{CEM}(\;\bullet\!\!-\!\!\bullet\;) &= \;\bullet\!\!-\!\!\bullet\;\otimes\bullet+\bullet^2\otimes\;\bullet\!\!-\!\!\bullet\;,\\
\Delta_{CEM}(\;\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\;) &= \;\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\;\otimes\bullet+2\bullet\;\bullet\!\!-\!\!\bullet\;\otimes\;\bullet\!\!-\!\!\bullet\;+\bullet^3\otimes\;\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\;,\\
\Delta_{CEM}(\;\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\;) &= \;\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\;\otimes\bullet+2\bullet\;\bullet\!\!-\!\!\bullet\;\otimes\;\bullet\!\!-\!\!\bullet\;+\bullet^3\otimes\;\bullet\!\!-\!\!\bullet\!\!-\!\!\bullet\;.
\end{aligned}\tag{1.21}$$

As shown in [CEFM09], it gives rise to a new Hopf algebra of trees, interacting with the Hopf algebra of trees of Connes and Kreimer.

1.2 High weak order stochastic integrators based on modified equations

We explain in this section how the approach of numerical integrators for modified equations introduced in [CHV07b] and described in Sect. 1.1.2 can be generalized to stochastic differential equations. This is a summary of the work [ACVZ12] in collaboration with A. Abdulle, D.Cohen, and K.C. Zygalakis.

1.2.1 Weak stochastic integrators

We consider a general Itô stochastic system of ordinary differential equations

$$dX(t) = f(X(t))dt + \sum_{r=1}^{m} g^r(X(t))dW_r(t), \quad X(0) = X_0, \tag{1.22}$$

where $X(t)$ is a random variable with values in $\mathbb{R}^N$, $f : \mathbb{R}^N \to \mathbb{R}^N$ is the drift term, $g^r : \mathbb{R}^N \to \mathbb{R}^N$, $r = 1, \dots, m$ are the diffusion terms, and $W_r(t)$, $r = 1, \dots, m$ are independent one-dimensional Wiener processes. The drift and diffusion functions are assumed smooth enough, Lipschitz continuous and to satisfy a growth bound in order to ensure a unique (mean-square bounded) solution of (1.22) [Arn74, KP92]. For the numerical approximation of (1.22) we consider the discrete map

$$X_{n+1} = \Psi(X_n, h, \xi_n), \tag{1.23}$$

where $\Psi(\cdot, h, \xi_n) : \mathbb{R}^N \to \mathbb{R}^N$, $X_n \in \mathbb{R}^N$ for $n \geq 0$, h denotes the timestep size, and ξ_n denotes a random vector. The numerical approximation (1.23), starting from the exact initial condition X_0 of (1.22) is said to have weak order τ if for all functions[2] $\phi : \mathbb{R}^N \to \mathbb{R} \in C_P^{2(\tau+1)}(\mathbb{R}^N, \mathbb{R})$,

$$|\mathbb{E}(\phi(X_n)) - \mathbb{E}(\phi(X(t_n)))| \leq Ch^\tau, \tag{1.24}$$

and to have strong order τ if

$$\mathbb{E}(|X_n - X(t_n)|) \leq Ch^\tau, \tag{1.25}$$

for any $t_n = nh \in [0, T]$ with $T > 0$ fixed, for all h small enough, with constants C independent of h.

[2] Here and in what follows, $C_P^\ell(\mathbb{R}^N, \mathbb{R})$ denotes the space of ℓ times continuously differentiable functions $\mathbb{R}^N \to \mathbb{R}$ with all partial derivatives with polynomial growth.

Remark 1.2.1 *A well-known theorem of Milstein [Mil86] (see [MT04, Chap. 2.2]) allows to infer the global weak order from the error after one step. Assuming that $f, g^r \in C_P^{2(\tau+1)}(\mathbb{R}^N, \mathbb{R}^N), r = 1, \ldots, m$ are Lipschitz continuous, that for all $r \in \mathbb{N}$, the moments $\mathbb{E}(|X_n|^{2r})$ are bounded for all n, h with $0 \leq nh \leq T$ uniformly with respect to all h sufficiently small, and that the local error bound for all $\phi \in C_P^{2(\tau+1)}(\mathbb{R}^N, \mathbb{R})$ and all initial values $X(0) = X_0$ satisfies*

$$|\mathbb{E}(\phi(X_1)) - \mathbb{E}(\phi(X(t_1)))| \leq Ch^{\tau+1} \tag{1.26}$$

for all h sufficiently small, then the global error bound (1.24) *holds. Here the constant C is again independent of h. For the strong convergence we have the following result [Mil87]. If the functions f, g^r are sufficiently smooth and Lipschitz continuous and*

$$\mathbb{E}|X_1 - X(t_1)| \leq Ch^{\tau+1/2} \quad \textit{and} \quad |\mathbb{E}(X_1) - \mathbb{E}(X(t_1))| \leq Ch^{\tau+1}, \tag{1.27}$$

for all initial values $X(0) = X_0$, then the global error bound (1.25) *holds.*

The simplest method to approximate solutions to the Itô SDE (1.22) is the so-called Euler-Maruyama method

$$X_{n+1} = X_n + hf(X_n) + \sum_{r=1}^{m} g^r(X_n)\Delta W_{n,r}, \tag{1.28}$$

where $\Delta W_{n,r} \sim \mathcal{N}(0, h),\ r = 1, \ldots m$ are independent Wiener increments. This method has strong order $1/2$ and weak order 1 for a general system of Itô SDEs [Mar55]. Various higher order weak methods have been considered in the literature [KP92, MT04]. For example, weak second order methods were proposed by Milstein [Mil78, Mil86], Platen [Pla92], Talay [Tal84] and Tocino and Vigo-Aguiar [TVA02], and more recently classes of Runge-Kutta type methods were proposed by Rößler [Rößl03]. We mention also the extrapolation methods of Talay and Tubaro [TT90] and of [KPH95] that combines methods with different stepsizes to achieve higher weak order convergence. Higher order integrators can be constructed. We mention the classical Milstein-Talay method [Tal84] which has strong order one and weak order two,

$$\begin{aligned} X_1 &= X_0 + hf(X_0) + \sum_{r=1}^{m} g^r(X_0)\Delta W_r + \sum_{q,r=1}^{m} (g^r)'(X_0)g^q(X_0)I_{q,r} \\ &+ \frac{h^2}{2}\Big(f'(X_0)f(X_0) + \frac{1}{2}\sum_{r=1}^{m} f''(X_0)(g^r(X_0), g^r(X_0))\Big) + \sum_{r=1}^{m} f'(X_0)g^r(X_0)I_{r,0} \\ &+ \sum_{r=1}^{m}\Big((g^r)'(X_0)f(X_0) + \sum_{q=1}^{m}\frac{1}{2}(g^r)''(X_0)(g^q(X_0), g^q(X_0))\Big)I_{0,r}, \end{aligned} \tag{1.29}$$

where $I_{r,0}, I_{0,r}, I_{q,r}$ denote the stochastic integrals defined by

$$I_{r,0} = \int_{t_0}^{t_1}\int_{t_0}^{t} dW_r(s)dt,\ I_{0,r} = \int_{t_0}^{t_1}\int_{t_0}^{t} ds\ dW_r(t),\ I_{q,r} = \int_{t_0}^{t_1}\int_{t_0}^{t} dW_q(s)dW_r(t). \tag{1.30}$$

For notational brevity, we shall always write X_1 and X_0 in place of X_{n+1} and X_n when introducing an integrator.

Remark 1.2.2 *As given above, the method* (1.29) *is not practical for implementation: it contains derivatives which are expensive in general, and stochastic integrals that are difficult to simulate. If one is only interested in the weak convergence, a standard approach is to replace these stochastic multiple integrales by appropriate weak approximation with discrete random variable. This will discussed and exploited in Section 2.2.*

1.2.2 Modified stochastic differential equations

The general idea of constructing high order integrators based on modifying equation can be generalized to SDEs as follows. Consider a numerical method (1.23) for problem (1.22), with smooth vector fields $f, g^r, r = 1, \ldots, m$, and assume that its weak order of convergence (1.24) is $p \geq 1$. We show that under suitable assumptions, the weak order p of the numerical integrator (1.23) can be increased to $p+r$ with $r \geq 1$ by applying it to a suitably modified SDE

$$d\widetilde{X} = f_h(\widetilde{X})dt + g_h(\widetilde{X})dW(t), \quad \widetilde{X}(0) = X_0, \tag{1.31}$$

with modified drift and noise of the form

$$f_h(x) = f(x) + hf_1(x) + \ldots + h^s f_s(x), \tag{1.32}$$
$$g_h(x) = g(x) + hg_1(x) + \ldots + h^s g_s(x), \tag{1.33}$$

where $s = p + r - 1$. The integrator with improved weak order r can be written as

$$\widetilde{X}_{n+1} = \Psi(f_{h,p+r-1}, g_{h,p+r-1}, \widetilde{X}_n, h, \xi_n). \tag{1.34}$$

Remark 1.2.3 *The above procedure should not be confused with a procedure called backward error analysis for SDEs [DF12, Zyg11] or the related approach [Sha06], developed to study the long time behavior of numerical methods for SDEs. There, one tries to find a modified equation*

$$d\hat{X} = a_h(\hat{X})dt + b_h(\hat{X})dW(t), \quad \hat{X}(0) = X_0, \tag{1.35}$$

such that its exact solution is closer to the numerical solution (1.23)*, i.e.,*

$$|\mathbb{E}(\phi(X_N)) - \mathbb{E}(\phi(\hat{X}(t_N)))| \leq Ch^{p+q},$$

with $q > 0$. In general, the modified SDEs (1.35) *and* (1.31) *are different similarly to the deterministic case.*

1.2.3 Construction of high-order integrators based on modified equations

A natural and standard way of looking at expectations of functionals of diffusion processes is by using the backward Kolmogorov equation associated to (1.22), which is the (deterministic) partial differential equation

$$\frac{\partial u}{\partial t} = \mathcal{L}u, \quad u(x, 0) = \phi(x), \tag{1.36}$$

where $\phi : \mathbb{R}^d \to \mathbb{R}$ is a smooth function, and the differential operator $\mathcal{L}$, called the generator of the SDE, is given by

$$\mathcal{L} := f \cdot \nabla_x + \frac{1}{2}(gg^T) : \nabla_x^2. \tag{1.37}$$

In (1.37), ∇_x and ∇_x^2 denote respectively the gradient and the Hessian matrix operator [3] with respect to x. In the case $m = d = 1$, the generator reduces to

$$\mathcal{L} = f\frac{\partial}{\partial x} + \frac{1}{2}g^2\frac{\partial^2}{\partial x^2}.$$

[3] Here, we consider the usual scalar product on matrices defined by $A : B = \text{trace}(A^T B)$.

The probabilistic interpretation (see for example [Øks03, PS08, Ris89]) of the solution $u = u^{f,g}(\phi, x, t)$ to (1.36) is that

$$u^{f,g}(\phi, x, t) = \mathbb{E}\left(\phi(X(t))|X(0) = x\right),$$

where $X(t)$ solves (1.22). Using (1.36) one can easily derive the following formal Taylor expansion [DF12, Zyg11]

$$u^{f,g}(\phi, x, h) - \phi(x) = \sum_{j=1}^{\infty} \frac{h^j}{j!} \mathcal{L}^j \phi(x).$$

Under appropriate smoothness assumptions on f, g and ϕ one can prove that

$$u^{f,g}(\phi, x, h) - \phi(x) = \sum_{j=1}^{k} \frac{h^j}{j!} \mathcal{L}^j \phi(x) + \mathcal{O}(h^{k+1}), \tag{1.38}$$

for all integer k. By defining

$$U^{f,g}(\phi, x, h) = \mathbb{E}\left(\phi\left(\Psi(f, g, X_0, h, \xi_0)|X_0 = x\right)\right), \tag{1.39}$$

for the numerical integrator (1.23), we see that the local weak error of the numerical integrator applied to (1.22) after one step is given by

$$\mathbb{E}(\phi(X_1)) - \mathbb{E}(\phi(X(t_1))) = U^{f,g}(\phi, x, h) - u^{f,g}(\phi, x, h). \tag{1.40}$$

Note that (1.40) is the reformulation of the left-hand side of the local error bound (1.26) in terms of the solution of the backward Kolmogorov equation (1.36) associated to (1.22). Motivated by an expansion of (1.39) in Taylor series, we assume

Assumption 1.2.4 *The numerical solution* (1.39) *has the following expansion*

$$U^{f,g}(\phi, x, h) = \phi(x) + hA_0(f, g)\phi(x) + h^2 A_1(f, g)\phi(x) + \ldots, \tag{1.41}$$

where $A_i(f, g),\ i = 0, 1, 2, \ldots$ are differential operators depending on the drift and diffusion functions of the SDE to which the numerical integrator is applied to. We further assume that these differential operators $A_i(f, g),\ i = 0, 1, 2, \ldots$ satisfy for all $f, \hat{f}, g, \hat{g}$ and $\varepsilon \to 0$,

$$A_i(f + \varepsilon\hat{f}, g + \varepsilon\hat{g}) = A_i(f, g) + \varepsilon\hat{A}_i(f, \hat{f}, g, \hat{g}) + \mathcal{O}(\varepsilon^2),$$

where $\hat{A}_i(f, \hat{f}, g, \hat{g}),\ i = 0, 1, 2, \ldots$ are again differential operators.

The above smoothness hypothesis is usually satisfied by numerical integrators. For the assumption that the expansion has integer powers of the stepsize h, special care has to be taken (see [ACVZ12, Rem. 2.2] for details).

Theorem 1.2.5 *Assume that the numerical scheme* (1.23) *has order $p \geq 1$ and that Assumption 1.2.4 holds. Let $r \geq 1$ and assume that the functions f_j and g_j for $j = 1, \ldots, p+r-2$ have been constructed such that $\widetilde{X}_{n+1} = \Psi(f_{h,p+r-2}, g_{h,p+r-2}, \widetilde{X}_n, h, \xi_n)$ has weak order $p + r - 1$. Consider the differential operator defined as*

$$\mathcal{L}_{p+r-1} := \lim_{h\to 0} \frac{u^{f,g}(\cdot, x, h) - U^{f_{h,p+r-1}, g_{h,p+r-1}}(\cdot, x, h)}{h^{p+r}}, \tag{1.42}$$

where $u^{f,g}(\phi,x,h)$ *is expanded in* (1.38) *and* $U^{f,g}(\phi,x,h)$ *is defined in* (1.39). *If there exist functions* $f_{p+r-1}:\mathbb{R}^d\to\mathbb{R}^d$ *and* $g_{p+r-1}:\mathbb{R}^d\to\mathbb{R}^{d\times m}$ *such that the differential operator* (1.42) *can be written in the form* $\mathcal{L}_j = f_j\cdot\nabla_x + \frac{1}{2}\sum_{k=0}^{j}(g_k g_{j-k}^T):\nabla_x^2$, *(where* $f_0 := f$ *and* $g_0 := g$*), then the numerical integrator* (1.34) *applied to the SDE with the modified drift and noise* (1.32),(1.33) *has weak order of accuracy* $p+r$ *for the original system of SDEs* (1.22) *provided* $f_{h,p+r-1}, g_{h,p+r-1} \in C_P^{2(p+r)+1}(\mathbb{R}^d,\mathbb{R}^d)$. *The error bound*

$$|\mathbb{E}(\phi(\widetilde{X}_N)) - \mathbb{E}(\phi(X(t_N)))| \le Ch^{p+r},$$

holds for any fixed $t_N = Nh \in [0,\tau]$ *with* h *sufficiently small and for all functions* $\phi \in C_P^{2(p+r)+1}(\mathbb{R}^d,\mathbb{R})$.

1.2.4 Application: the modified θ-Milstein method

Consider the multi-dimensional SDE (1.22), where $f, g^r, r=1,\dots,m$ are (smooth) column vector fields of size d. For a fixed parameter θ, consider the θ-Milstein method, which has strong and weak orders one,

$$X_1 = X_0 + (1-\theta)hf(X_0) + \theta hf(X_1) + \sum_{r=1}^{m} g^r(X_0)\Delta W_0 + M(X_0,W), \tag{1.43}$$

where the Milstein term $M(X_0,W)$ is defined by

$$M(X_0,W) = \sum_{q,r=1}^{m} (g^r)'(X_0)g^q(X_0)I_{q,r}$$

where $I_{q,r}$ is defined in (1.30). Applying the modified equation approach to increase the weak order yields the modified θ-Milstein method of weak order two

$$X_1 = X_0 + (1-\theta)hf(X_0) + \theta hf_{h,1}(X_1) + \sum_{r=1}^{m} g_{h,1}^r(X_0)\Delta W_0 + M(X_0,W), \tag{1.44}$$

where $f_{h,1} = f + hf_1$, $g_{h,1}^r = g^r + hg_1^r$, and $f_1, g_1^r, r=1,\dots,m$ are given by

$$\begin{aligned} f_1 &= \left(\frac{1}{2}-\theta\right)(f'f) + \frac{1}{2}\left(\frac{1}{2}-\theta\right)\sum_{r=1}^{m} f''(g^r,g^r), \\ g_1^r &= \left(\frac{1}{2}-\theta\right)f'g^r + \frac{1}{2}(g^r)'f + \frac{1}{4}\sum_{r=1}^{m} g''(g^r,g^r), \end{aligned} \tag{1.45}$$

for all $i=1,\dots,d$ and $j=1,\dots,m$.

Remark 1.2.6 *It can be shown that the functions* g^r *in the definition of* $M(X_0,W)$ *can remain unchanged without affecting the weak order two of the modified* θ*-Milstein method* (1.44).

Remark 1.2.7 *It can be shown that the modified* θ*-Milstein method has the mean-square A-stable property (a standard notion discussed further in Section 2.1) for* $\theta = 1$. *This makes this integrator suitable for the numerical integration of stiff systems of SDEs. Note that the integrator* (1.44) *belongs to a sub-class of a general family of weak second order methods introduced by Milstein [Mil86]. For* $\theta = 0$ *it has also been considered by Talay who proved its order of convergence [Tal84]. For* $\theta = 1/2$ *the method was considered by Milstein who showed its good stability behavior for scalar SDEs with additive noise. For* $\theta = 1$, *the method does not seem to have appeared explicitly in the literature.*

We observe that the above method contains derivatives of the drift and diffusion functions. This is a general feature of the methods obtained using modified equations. In some cases, these derivatives are easy and cheap to compute (see for example the stochastic mechanical problem in the next Section 1.2.5). In general, these derivatives can be approximated. In particular, one can use formulas based on finite differences. Some care is however required for an efficient implementation (i.e., a low number of function evaluations in dependence on the number of Wiener processes [DR09b]).

1.2.5 Application: the modified stochastic implicit midpoint rule

Another application of the modified equation strategy is the construction of numerical integrators for Stratonovich SDEs of high weak order which exactly conserve all quadratic first integrals (up to machine precision). We consider the SDE (1.22) in Stratonovich form with a one-dimensional noise

$$dX = f(X)dt + g(X) \circ dW(t), \quad X(0) = X_0, \tag{1.46}$$

where the notation $\circ dW(t)$ emphasizes that the Stratonovich stochastic integrals are considered for (1.46). As a basic numerical integrator to apply our methodology of modified equation, we choose the (fully) implicit midpoint rule, as first introduced in [MRT02],

$$X_{n+1} = X_n + hf\left(\frac{X_n + X_{n+1}}{2}\right) + g\left(\frac{X_n + X_{n+1}}{2}\right)\Delta W_n, \tag{1.47}$$

where ΔW_n is a scalar random variable. It is shown in [MRT02] that (1.47) has weak and strong orders one in the case of a one-dimensional or commutative multi-dimensional noise. Notice however that for general SDEs with multi-dimensional noise, the strong order is $1/2$ and the weak order is 1.

First integral conservation A smooth quantity $C(x)$ is called a first integral of the system (1.46) if it is exactly conserved along time for all realizations of the Wiener process $W(t)$, i.e. $C(X(t)) = C(X_0)$ for all time t and all initial condition $X(0) = X_0$. Given a smooth function $C(x)$, the identity[4] $dC(X) = \nabla C(X) \cdot f(X)dt + \nabla C(X) \cdot g(X) \circ dW(t)$ shows that $C(X)$ is a first integral of (1.46) if and only if

$$\nabla C(x) \cdot f(x) = \nabla C(x) \cdot g(x) = 0 \quad \text{for all } x \in \mathbb{R}^d. \tag{1.48}$$

Remark 1.2.8 *The method* (1.47) *is implicit with respect to both the drift and the noise terms. In the case where ΔW_n is a standard Gaussian variable, the unboundedness of ΔW_n for arbitrarily small h leads to non-uniqueness of solutions to the non-linear system* (1.47) *and the integrator is not well defined. One way to address this problem, is to replace ΔW_n, with a suitable chosen bounded random variable [MRT02] (see also [MT04, Sect. 1.3]). Here we shall simply consider three point discrete random variable (see χ_r in* (2.10) *in Section 2.2), which are obviously bounded.*

Using the framework of integrators based on modified equations, we introduce the following new numerical integrator of weak second order for the SDE (1.46) which preserves all quadratic first integrals.

[4] Note that Stratonovich calculus is used here.

Algorithm 1.2.9 (Modified stochastic midpoint rule of weak second order)

$$X_{n+1} = X_n + hf_{h,1}\left(\frac{X_n + X_{n+1}}{2}\right) + g_{h,1}\left(\frac{X_n + X_{n+1}}{2}\right)\Delta W_n, \tag{1.49}$$

where $f_{h,1} = f + hf_1$ *and* $g_{h,1} = g + hg_1$ *and*

$$f_1 = \frac{1}{4}\left(\frac{1}{2}f''(g,g) - g'f'g\right) \qquad g_1 = \frac{1}{4}\left(\frac{1}{2}g''(g,g) - g'g'g\right). \tag{1.50}$$

We obtain that if we consider the modified Stratonovich SDE

$$dX = [f(X) + hf_1(X)]\,dt + [g(X) + hg_1(X)] \circ dW(t), \tag{1.51}$$

then (1.49) is equivalent to applying the original midpoint rule (1.47) to the modified Stratonovich SDE (1.51).

Theorem 1.2.10 *The integrator* (1.49) *for a system of Stratonovich SDEs* (1.46) *with* $m = 1$ *noise has weak order* 2. *It exactly conserves all quadratic first integrals of* (1.46).

Example: a stochastic rigid body model To illustrate that the integrators previously introduced conserve quadratic first integrals and to compare the performance of the proposed high-order integrators preserving quadratic first integrals, we consider a randomly perturbed rigid body problem that is, the motion of a rigid body in $\mathbb{R}^3$ subject to a scalar white noise perturbation. The equations of motion of an asymmetric rigid body with Stratonovich noise in dimension $m = 1$ are given by [5]

$$\begin{aligned} dy &= \widehat{y}\mathcal{I}^{-1}y\,dt + \mu\widehat{y}e_1 \circ dW(t), \\ dQ &= Q\widehat{\mathcal{I}^{-1}y}\,dt + \mu Q\widehat{e_1} \circ dW(t), \end{aligned} \tag{1.52}$$

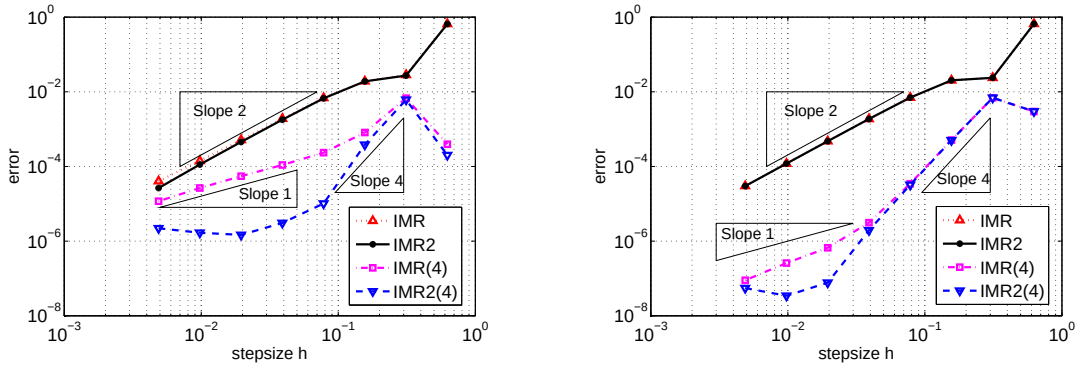
where $e_1 = (1,0,0)^T$, $\mu \geq 0$ is a parameter and $\mathcal{I} = \mathrm{diag}\,(I_1, I_2, I_3)$. A generalization of equation (1.52) for a 3-dimensional noise is presented in [LCO09, Eq. (6.9)-(6.10)], where one can also find a physical justification for these equations. This model is a variant of the model proposed in [Lia97] with the additional feature that it preserves the spatial angular momentum Qy. In the case where $\mu = 0$, we recover the standard deterministic equations of motion of an asymmetric rigid body (1.8). The system of SDEs (1.52) has the same first integrals as in the deterministic case (all of which are quadratic), with the exception of the Hamiltonian in (1.11). Indeed, $dH(y) = \mu\frac{y_{[2]}y_{[3]}}{2}\left(\frac{1}{I_2} - \frac{1}{I_3}\right) \circ dW(t)$, is non zero in general (unless $\mu(I_2 - I_3) = 0$).

Using formulas (1.50) where the functions f and g correspond to the right-hand side of (1.52), a straightforward computation yields the modified SDE associated to (1.52),

$$\begin{aligned} dy &= \widehat{y}(\mathcal{I}^{-1} + \frac{h\mu^2}{4}\mathcal{J}^{-1})y\,dt + \mu\left(1 + \frac{h\mu^2}{4}\right)\widehat{y}e_1 \circ dW(t), \\ dQ &= Q(\widehat{\mathcal{I}^{-1}y} + \frac{h\mu^2}{4}\widehat{\mathcal{J}^{-1}y})dt + \mu\left(1 + \frac{h\mu^2}{4}\right)Q\widehat{e_1} \circ dW(t), \end{aligned} \tag{1.53}$$

where we define $\mathcal{J} = \mathrm{diag}\,(I_1, I_3, I_2)$. We obtain from Theorem 1.2.10 that applying the implicit midpoint rule (1.47) to the Statonovitch SDE (1.53) yields a weak order two approximation of the solution of (1.52) which exactly conserves all quadratic first integrals,

[5] We use again the standard hat notation defined in (1.8).



(a) Error for $\mathbb{E}(\sin(e_1^T Q e_1))$. Noise size $\mu = 0.1$.

(b) Error for $\mathbb{E}(\sin(e_1^T Q e_1))$. Noise size $\mu = 0.01$.

Figure 1.2: Rigid body problem (1.52). Comparison of weak convergence rates for IMR, see (1.47) (dotted lines), IMR2, see (1.49) (solid lines), IMR2(4) (dashed lines), and IMR(4) (dashed-dotted lines).

i.e. $C(y_{n+1}) = C(y_n)$, $Q_{n+1}y_{n+1} = Q_n y_n$ and $Q_n^T Q_n = Id$ for all n, and in the case $I_2 = I_3$ (symmetric body), we have also $H(y_{n+1}) = H(y_n)$, where $C(y)$ and $H(y)$ are defined in (1.11).

Note that the modified SDE (1.53) is of the same form as the original equations (1.52) with modified data parameters $\widetilde{\mu} = \mu(1+h\mu^2/4)$,, $\widetilde{I}_1, \widetilde{I}_2, \widetilde{I}_3$. Thus, our modification to high weak order reduces to a perturbation of the parameters and has a negligible overcost.

Weak convergence rates In Figure 1.2, we compare the performance of the modified stochastic implicit midpoint rule (denoted IMR2) applied to the stochastic rigid body model (1.52) to the standard implicit midpoint rule (1.47) (denoted IMR). For comparison, we also include the methods IMR(4) and IMR2(4) which are the same methods as IMR and IMR2 with the exception that the deterministic $\mathcal{O}(h^2)$ correction term from (1.7) is also included. We take the moments of inertia $I_1 = 0.345$, $I_2 = 0.653$, $I_3 = 1.0$, which correspond to the water molecule (nearly flat body). Initial values are $X(0) = (0.8, 0.6, 0)^T$ and $Q(0)$ is the identity matrix. We plot the errors for $\mathbb{E}(\sin(e_1^T Q e_1))$ at final time $t = 10$ versus the timestep $h = 2^{-i}$, $i = 1, \ldots, 8$. The reference solution is computed using the small timestep $h = 2^{-14}$. To check carefully the accuracy of the methods, we compute numerically $\mathbb{E}(\sin(e_1^T Q e_1))$ using the averages over 300 millions of trajectories (this computation was performed using the EPFL cluster on one hundred independent CPUs). We consider two values of the noise parameter: $\mu = 0.1$ and $\mu = 0.01$. We observe in all cases lines of slope two for the modified midpoint rule IMR2 (1.49) which confirms its weak order two of accuracy. For the standard midpoint rule IMR in (1.47) and the modified version IMR(4) which both have weak order one, we observe for large stepsizes h, lines of slope four and two respectively in the case where the deterministic error (h^2 or h^4) is dominant compared to the stochastic error with size $\mu^2 h$.

1.3 Multi-revolution composition methods for highly oscillatory problems

This section summarizes the work [CMMV13]. The aim is to construct geometric integrators for highly oscillatory problems allowing the use of large timesteps, with uniform

accuracy with respect to the highly oscillatory frequency of the problem.

Although this work cannot be cast directly in the framework of integrators based on modified equations, the idea is to capture an averaged system's solution (in the spirit of homogenization methods) with appropriate modifications of the oscillatory period.

The originality of the approach is to apply geometric integration techniques (high-order composition methods) originally designed for non-stiff problems to highly oscillatory problems which are stiff and for which the computational cost of standard integrators grows with the stiffness due to accuracy and stability constraints. Indeed, for standard integrators stability and accuracy requirements induce a step-size restriction of the form $h \leq C\varepsilon$, where $1/\varepsilon$ is the highly oscillatory frequency, which renders the computation of a reasonably accurate solution more and more costly and sometimes even untractable for small values of ε.

Approximating iterations of a near identity map We are concerned with the approximation of the M-th iterates of a near-identity smooth map by compositions methods. More precisely, considering a smooth map $(\varepsilon, y) \mapsto \varphi_\varepsilon(y)$ of the form

$$\varphi_\varepsilon(y) = y + \varepsilon\Theta_\varepsilon(y), \tag{1.54}$$

we wish to approximate the result of $M = \mathcal{O}(1/\varepsilon)$ compositions of φ_ε with itself

$$\varphi_\varepsilon^M = \underbrace{\varphi_\varepsilon \circ \cdots \circ \varphi_\varepsilon}_{M \text{ times}} \tag{1.55}$$

with the aid of a method whose efficiency remains essentially independent of ε.

In order to motivate our composition methods, it will be usefull to observe that φ_ε can be seen as one step with step-size ε of a first order integrator for the differential equation

$$\frac{dz(t)}{dt} = \Theta_0(z(t)), \tag{1.56}$$

where $\Theta_0(z) = \frac{d}{d\varepsilon}\varphi_\varepsilon(z)\big|_{\varepsilon=0}$, and thus, $\varphi_\varepsilon^M(y)$ may be interpreted as an approximation at $t = M\varepsilon$ of the solution $z(t)$ of (1.56) with initial condition

$$z(0) = y. \tag{1.57}$$

A standard error analysis shows that $\varphi_\varepsilon^N(y) - z(N\varepsilon) = \mathcal{O}(\varepsilon H)$ as $H = N\varepsilon \to 0$, which makes clear that, for sufficiently small $H = \varepsilon N$, $\varphi_\varepsilon^N(y)$ could be approximated by one step $\Psi_H(y) \approx z(H)$ of any pth order integrator applied to the initial value problem (1.56)–(1.57) within an error of size $\mathcal{O}(H^{p+1} + \varepsilon H)$. In particular, φ_H can be seen as a first order integrator for the ODE (1.56), and a second order integrator can be obtained as

$$\Psi_H(y) = \varphi_{H/2} \circ \varphi_{H/2}^*(y), \tag{1.58}$$

where $\varphi_\varepsilon^* := \varphi_{-\varepsilon}^{-1}$ is the *adjoint map* of φ_ε.

New class of multi-revolution composition methods Motivated by that, we generalize the above approximation by considering coefficients $\alpha_j, \beta_j, j = 1, \ldots$ depending on N, and chosen in such a way that φ_ε^N is approximated for sufficiently small $H = N\varepsilon$ within an error of size $\mathcal{O}(H^{p+1})$, where the error constant is independent of N, H, ε. We say that the composition method

$$\Psi_{N,H}(y) := \varphi_{\alpha_1(N)H} \circ \varphi_{\beta_1(N)H}^* \circ \cdots \circ \varphi_{\alpha_s(N)H} \circ \varphi_{\beta_s(N)H}^*(y) \tag{1.59}$$

is an s-stage pth order *multi-revolution composition method* (MRCM) if

$$\Psi_{N,H}(y) = \varphi_\varepsilon^N(y) + \mathcal{O}(H^{p+1}), \quad \text{for} \quad H = N\varepsilon \le H_0. \tag{1.60}$$

For instance, we will see that the second order standard composition method (1.58) can be modified to give a second order MRCM (1.59) with $s = 1$, $\alpha_1(N) = (1 + N^{-1})/2$, and $\beta_1(N) = (1 - N^{-1})/2$,

$$\Psi_{N,H}(y) = \varphi_{\alpha_1(N)H} \circ \varphi^*_{\beta_1(N)H}(y) = \varphi_\varepsilon^N(y) + \mathcal{O}(H^3), \quad H = N\varepsilon.$$

It is interesting to observe that this second order MRCM reduces in the limit case $N \to \infty$ to the standard composition method (1.58) (a second order integrator for the ODE (1.56)), which is consistent with the fact that $\varphi_{H/N}^N$ converges to the H-flow of (1.56) as $N \to \infty$. More generally, any pth order MRCM (1.59), gives rise to a pth order standard composition method with coefficients

$$a_i = \lim_{N\to\infty} \alpha_i(N), \quad b_i = \lim_{N\to\infty} \beta_i(N).$$

In practice, if one wants to approximately compute the map φ_ε^M for a given small value of ε and large positive integers M within a given error tolerance by means of a s-stage pth order MRCM (1.59), then one should choose a sufficiently small step-size H to achieve the required accuracy, and accordingly choose N as the integer part of H/ε, in order to approximate $\varphi_\varepsilon^M(y)$, for $M = mN$, $m = 1, 2, 3 \ldots$, as $\varphi_\varepsilon^{mN}(y) \approx \Psi_{N,H}(y)^m$.

Application to highly oscillatory problems The main application we have in mind is the time integration of highly-oscillatory problems with a single harmonic frequency $\omega = 2\pi/\varepsilon$. In the numerical examples, we consider in particular problems of the form

$$\frac{d}{dt}y(t) = \frac{1}{\varepsilon}Ay(t) + f(y(t)), \qquad 0 \le t \le T, \qquad y(0) = y_0 \in \mathbb{R}^d, \tag{1.61}$$

where A is a $d \times d$ skew-symmetric matrix with eigenvalues in $2\pi i\mathbb{Z}$, so that e^{tA} is 1-periodic in time, and where $f : \mathbb{R}^d \to \mathbb{R}^d$ is a given nonlinear smooth function. In this situation, we shall consider φ_ε as the flow with time ε (the period of the unperturbed equation corresponding to $f(y) \equiv 0$) of equation (1.61), or equivalently, the flow with time 1 of the system

$$\frac{d}{dt}y(t) = Ay(t) + \varepsilon f(y(t)).$$

It is well known [CMSS10, CMSS12] that such a map φ_ε is a smooth near-identity map, and furthermore, that (1.56) is in this case the first order averaged equation, more precisely,

$$\Theta_0(z) = \frac{d}{d\varepsilon}\varphi_\varepsilon(z)\Big|_{\varepsilon=0} = \int_0^1 e^{-At} f(e^{At}z)dt.$$

The solution $y(t)$ of the initial value problem (1.61) sampled at the times $t = \varepsilon M$ will then be given by

$$y(\varepsilon M) = \varphi_\varepsilon^M(y_0),$$

and thus, for an appropriately chosen positive integer N (determined by accuracy requirements and the actual value of ε), we may use a pth order MRCM (1.59) to compute the approximations

$$y_m = \Psi_{N,H}(y)^m \approx \varphi_\varepsilon^{mN}(y_0) = y(t_m), \quad \text{where} \quad t_m = mH, \quad H = \varepsilon N.$$

The local error estimate (1.60) then leads by standard arguments to a global error estimate of the form

$$y_m - y(t_m) = \mathcal{O}(H^p), \quad \text{for} \quad t_m = mH \leq T,$$

where the constant in the $\mathcal{O}$-term depends on T but is independent of ε and H.

Derivation of general order conditions In [CMMV13], we derive general order conditions for multi-revolution composition methods (1.59) to satisfy (1.60) (see Table 1.1 for conditions up to order four). This is done by comparing the Taylor expansions of both

Order 1:	①	$\sum_{k=1}^{s}(\alpha_k + \beta_k) = 1$
Order 2:	②	$\sum_{k=1}^{s}(\alpha_k^2 - \beta_k^2) = N^{-1}$
Order 3:	③	$\sum_{k=1}^{s}(\alpha_k^3 + \beta_k^3) = N^{-2}$
	① — ②	$\sum_{k=1}^{s}(\alpha_k^2 - \beta_k^2)\sum_{\ell=1}^{k}{}'(\alpha_\ell + \beta_\ell) = \frac{N^{-1} - N^{-2}}{2}$
Order 4:	④	$\sum_{k=1}^{s}(\alpha_k^4 - \beta_k^4) = N^{-3}$
	① — ③	$\sum_{k=1}^{s}(\alpha_k^3 + \beta_k^3)\sum_{\ell=1}^{k}{}'(\alpha_\ell + \beta_\ell) = \frac{N^{-2} - N^{-3}}{2}$
	① ① — ②	$\sum_{k=1}^{s}(\alpha_k^2 - \beta_k^2)\Big(\sum_{\ell=1}^{k}{}'(\alpha_\ell + \beta_\ell)\Big)^2 = \frac{N^{-1}(1 - N^{-1})(2 - N^{-1})}{6}$

Table 1.1: Fourth-order conditions for MRCMs (1.59). The prime attached to a summation symbol indicates that the sum of α_ℓ^j is only from 1 to $k-1$ while the sum of β_ℓ^j remains for 1 to k

.

sides of $\Psi_{N,H}(y) \simeq \varphi_\varepsilon^N(y)$. Although conceptually easy, the task is rendered very intricate by the enormous number of terms and redundant order conditions naturally arising. For instance, for order 4, there are 21 order conditions, but 14 or them are superfluous, and the true number of independent order conditions is only 7. Explicit conditions for standard composition methods have been obtained in a systematic way in [MSS99] by using the formalism of B_∞-series and trees (a generalization of B-series presented in Section 1.1.3, involving trees with labelled vertices), and this is again the main tool in our context.

A micro-macro method Typically, the maps φ_μ and φ_μ^* in (1.59) with $\mu = \alpha_j(N)H$, $\mu = \beta_j(N)H$ $(j = 1, \ldots, s)$ can not be computed exactly. In the context of highly oscillatory systems, and in particular, for systems of the form (1.61), the actual (approximate) computation of φ_μ can be carried out essentially as a black-box operation: In practice, one may use any available implementation of some numerical integrator to approximate the flow with time 1 of the ODE

$$\frac{d}{dt}y(t) = Ay(t) + \mu f(y(t)). \tag{1.62}$$

In particular, φ_μ may be approximated by applying n steps of step-size $h = 1/n$ of an appropriate splitting method to (1.62), where n is chosen so as to resolve one oscillation. Let $\Phi_{\mu,h}(y)$ denote the approximation of φ_μ obtained in this way with a qth order splitting method, then the following estimate

$$\Phi_{\mu,h}(y) - \varphi_\mu(y) = \mathcal{O}(\mu^r h^q) \tag{1.63}$$

will be guaranteed to hold with $r = 1$. It is worth remarking that more refined estimates of the form $\mathcal{O}(\mu^{r_1} h^{q_1} + \ldots + \mu^{r_\ell} h^{q_\ell})$ can be obtained for certain splitting methods [McL95]. Observe that one can expect $r \geq 1$ in the right-hand side of (1.63) if (as in the case of splitting methods for (1.62),) $\Phi_{\mu,h}$ is constructed so that $\Phi_{0,h}(y) = \varphi_0(y) = y$. We next define the following fully-discrete MRCM,

$$\varphi_\varepsilon^N(y) \simeq \Psi_{N,H,h}(y) := \Phi_{\alpha_1 H,h} \circ \Phi_{-\beta_1 H,h}^{-1} \circ \cdots \circ \Phi_{\alpha_s H,h} \circ \Phi_{-\beta_s H,h}^{-1}(y) \tag{1.64}$$

in the spirit of Heterogenerous multiscale methods (see [AEEVE12, EE03, EEL$^+$07] and the work presented in Chapter 3) which combine the application of macro-steps of length H (to advance along the solution of (1.61)) with the application to (1.62) of some integrator with micro-steps of size $h = 1/n$ (where n is chosen large enough to resolve each oscillation).

Comparison with other multi-revolution integrators The general idea of multi-revolution methods has been first considered in astronomy, where ε-perturbation of periodic systems are recurrent, and named as such since these methods approximate many *revolutions* (N periods of time) by only a few (in our approach, $2s$ compositions then accounts for $2s$ revolutions with different values of the perturbation parameter ε). A class of multi-revolution Runge-Kutta type methods has then been studied in the context of oscillatory problems of the form (1.61) [CMR03, CMR07, CJMR04, MP97, PJY97]. Closely related methods where considered in [Kir88] and also in [CCMSS11].

Actually, MRCMs are *asymptotic preserving*, a notion introduced in the context of kinetic equations (see [Jin99], and the recent works [LM08, FJ10]) and ensuring that a method is uniformly accurate for a large range of values of the parameter ε with a computational cost essentially independent of ε.

The methods introduced here differ from existing other multi-revolution methods in that they are intrinsically geometric, since they solely use compositions of maps of the form φ_μ and φ_μ^{-1}, whose geometric properties are determined by equation (1.61). In particular, it is *symplectic* if (1.61) is *Hamiltonian*, *volume-preserving* if (1.61) is *divergence-free*, and shares the same invariants which are independent of ε as the flow of (1.61).

Application: the nonlinear Schrödinger equation Highly oscillatory problems of the form (1.61) are in particular obtained by appropriate discretization in space of several Hamiltonian partial differentiation equations, such as nonlinear versions of the wave equation and the Schrödinger equation. In Figure 1.3, we consider MRCMs for the following problem with nonlinear effects analyzed by B. Grébert and C. Villegas-Blas in [GVB11]. It consists of a nonlinear Schrödinger equation with a cubic nonlinearity $|u|^2 u$ multiplied by an excitation term of the form $2\cos(2x)$ and may be stated on the one-dimensional torus as

$$\begin{aligned} i\partial_t u &= -\Delta u + 2\varepsilon \cos(2x)|u|^2 u, \quad t \geq 0, \quad u(t,\cdot) \in H^s(\mathbb{T}_{2\pi}) \\ u(0,x) &= \cos x + \sin x. \end{aligned} \tag{1.65}$$

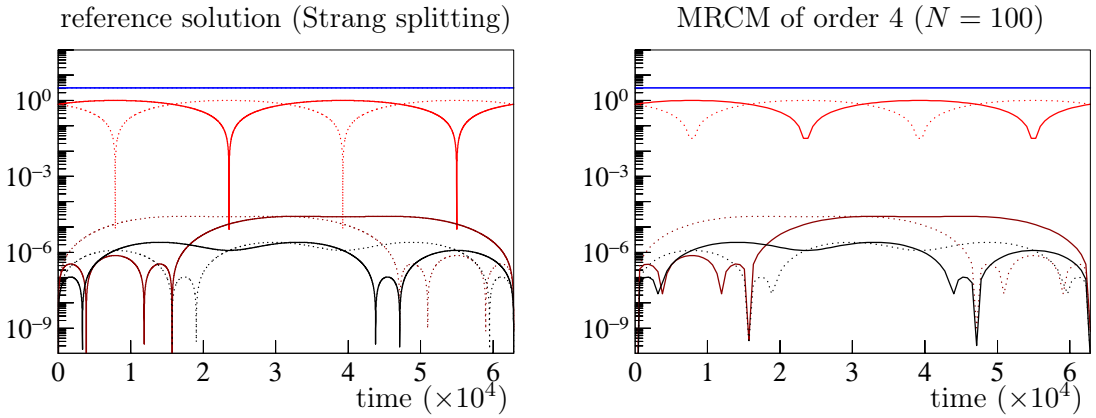


Figure 1.3: Nonlinear Schrödinger problem (1.65) with $\varepsilon = 10^{-4}$ on the time interval $(0, 2\pi\varepsilon^{-1})$. Plot of the actions $|\xi_j(t)|$, for $j = 1, 3, 5$ (solid lines) and for $j = -1, -3, -5$ (dotted lines) with colors red ($|j| = 1$), brown ($|j| = 3$), black ($|j| = 5$). The micro stepsize is $h = 2\pi/n$ with $n = 100$.

The problem is known to have a unique global solution in all Sobolev spaces $H^s(\mathbb{T}_{2\pi})$ for $s \geq 0$. A pseudospectral approximation of the form $u(t,x) \approx \sum_{k=-\ell}^{\ell} \xi_k(t) e^{ikx}$ may be obtained by determining the approximate Fourier modes $\xi_k(t)$ as the solution with appropriate initial values of a semidiscrete version of equation (1.65)

$$\frac{d}{dt}\xi_k = -ik^2\xi_k + \varepsilon\, f_k(\xi_{-\ell}, \ldots, \xi_{-1}, \xi_0, \xi_1, \ldots, \xi_\ell), \quad k = -\ell, \ldots, -1, 0, 1, \ldots, \ell. \tag{1.66}$$

Clearly, the system of ODEs (1.66) can be recast into the format (1.61) by rescaling time (that is, by rewriting the system in terms of the new time variable $\hat{t} = \frac{\varepsilon}{2\pi} t$). We plot in Figure 1.3 the solution (the action) obtained with a MRCM of order 4 with multirevolution parameter $N = 100$. We observe satisfactory solutions (with a characteristic nonlinear beating effect in the modes $\xi_{\pm 1}$) compared to the reference solution (Strang Splitting) at a reduced computational cost of about two orders of magnitude for $\varepsilon = 10^{-4}$.

1.4 Perspectives

There are two natural directions in which we would like to pursue our research in the context of modified differential equations.

- It is clear that the weak order p of a given integrator implies the same accuracy when sampling a given invariant measure. However, based on the framework of modified equations, it can be shown that this assumption can be relaxed (work [AVZ13a] in progress with A. Abdulle and K. Zygalakis), and simplified order conditions can be derived for the derivation of high order invariant measure sampling integrators.

- We would like also to extend the class of multi-revolution composition methods to a stochastic highly oscillatory context. There is a strong need in applications of efficient large time step integrators with favorable stability and geometric properties in particular for stochastic nonlinear Schrödinger equations. We would like also to explore the extension of splitting methods with complex times introduced in a deterministic setting [CCDV09] (reaction diffusion problems) for the construction of weak high order integrators with favorable geometric properties.

Chapter 2

Efficient implicit and explicit integrators for stiff stochastic problems

In Section 1.2, implicit weak second order methods with favorable geometric and/or stability properties were introduced using the framework [ACVZ12] of modified differential equations. This framework could in principle be used to construct higher order weak stabilized methods. Here we follow a different approach based on stabilizing a second weak order scheme originating from the weak second order Taylor method (1.29) known as the Milstein-Talay method [Tal84]. This chapter is organized as follows. In Section 2.1, we recall standard stability concepts that are essential for the understanding of stiff stochastic integrators. In Section 2.2 we describe an efficient implementation of the Milstein-Talay method (1.29) that will be useful to construct our stabilized integrators. We then construct successively a drift-implicit integrator (Section 2.3 summarizing [AVZ13b]) and an explicit stabilized integrator (Section 2.4, summarizing [AVZ12]) with extended mean-square stability domains. The developed methodology for the stabilization of stochastic integrators permits to construct a new "swiss-knife" integrator for diffusion problems with various stiffness regimes (advection, reaction, noise), see Section 2.5 summarizing [AV13b].

2.1 Standard stability concepts for SDEs

In practice it is not only the order of convergence that guarantees an efficient approximation of an SDE, but also the long-time behavior of the solution. Stability properties of the exact and the numerical solutions are important to understand this behavior. Widely used characterizations of stability for SDEs are the mean-square and the asymptotic stability (in the large) [Arn74, Has80]. The former measures the stability of moments, the latter measures the overall behavior of sample paths. In particular, we have the following definitions. The steady solution $X \equiv 0$ of a system of Itô SDEs (1.22) with $f(0) = g^r(0) = 0,\ r = 1, \dots, m$ is called stochastically asymptotically stable in the large if there exists $\delta > 0$ such that

$$\lim_{t\to\infty} |X(t)| = 0 \text{ with probability 1 for all } |X_0| < \delta, \tag{2.1}$$

mean-square stable if there exists $\delta > 0$, such that

$$\lim_{t\to\infty} \mathbb{E}(|X(t)|^2) = 0 \text{ for all } |X_0| < \delta. \tag{2.2}$$

2.1.1 The stochastic scalar test equation with multiplicative noise

To gain insight on the stability behavior of a numerical method, we consider a class of linear scalar test problems widely used in the literature [SM96, Hig00, BBT04, Toc05],

$$dX(t) = \lambda X(t)dt + \mu X(t)dW(t), \qquad X(0) = 1, \tag{2.3}$$

in dimensions $N = m = 1$, with fixed complex scalar parameters λ, μ. The exact solution of (2.3), given by $X(t) = \exp((\lambda - \frac{1}{2}\mu^2)t + \mu W(t))$, is stochastically asymptotically stable if and only if $\lim_{t\to\infty} |X(t)| = 0$ with probability 1, equivalently $(\lambda, \mu) \in \mathcal{S}^{AS}_{\rm SDE}$ with

$$\mathcal{S}^{AS}_{\rm SDE} := \big\{(\lambda, \mu) \in \mathbb{C}^2 \ ;\ \Re\big(\lambda - \tfrac{1}{2}\mu^2\big) < 0\big\}, \tag{2.4}$$

and mean-square stable if and only if $\lim_{t\to\infty} \mathbb{E}\big(|X(t)|^2\big) = 0$, equivalently $(\lambda, \mu) \in \mathcal{S}^{MS}_{\rm SDE}$ with

$$\mathcal{S}^{MS}_{\rm SDE} := \big\{(\lambda, \mu) \in \mathbb{C}^2 \ ;\ \Re(\lambda) + \tfrac{1}{2}|\mu|^2 < 0\big\}. \tag{2.5}$$

We name the domains $S^{MS}_{\rm SDE} \subset S^{AS}_{\rm SDE}$ the mean-square and asymptotic stability domains of the test equation (2.3), respectively.

Note that the justification of the test equation (2.3) is delicate for multi-dimensional systems. Already for multi-dimensional linear systems $dX = AXdt + \sum_{r=1}^{m} B_r X dW_r(t)$, where A, B_r are $N \times N$ matrices and dW_r are independent one-dimensional Wiener processes, it is difficult to extend the stability analysis of numerical integrators if A and $B_r,\ r = 1, \ldots, m$ do not commute and can thus not be simultaneously diagonalized. This has been investigated in [SM02, RB08] but these studies do not allow for an easy characterization of stability criteria. Using the theory of stochastic stabilization and destabilization [Mao94] an attempt to generalize the linear test equation has been proposed in [BK10], where two sets of test equations with $N = m = 2$ and $N = m = 3$ have been studied. The conclusion of these studies is that the stability behavior of the Euler-Maruyama method (or its generalization obtained by using the θ method for the drift term) is essentially captured by the test equation (2.3). We mention however that for linear systems with a non normal drift, the additional test equations in [BK10] capture stability behaviors (in particular in the pre asymptotic regime) of a numerical scheme that cannot be seen by studying (2.3). This phenomenon is also well known for systems of ODEs (see [HW96, IV.11]).

2.1.2 Stability of numerical integrators for SDEs

We now look for conditions such that a numerical method (1.23) applied to the linear test problem (2.3) yields numerically stable solutions. Similarly to the continuous case, we say that the numerical method (1.23) applied to (2.3) is said to be

- numerically asymptotically stable if $\lim_{n\to\infty} |X_n| = 0$ with probability 1;
- numerically mean-square stable if $\lim_{n\to\infty} \mathbb{E}(|X_n|^2) = 0$.

Applying a numerical method to the test SDE (2.3) usually yields [Hig00] the following one step difference equation

$$X_{n+1} = R(p, q, \xi_n) X_n, \tag{2.6}$$

where $p = \lambda h, q = \mu\sqrt{h}$, and ξ_n is a random variable (e.g. a Gaussian $\xi_n \sim \mathcal{N}(0, 1)$ or a discrete random variable). Once this difference equation is formulated, it is not difficult to define the domains of mean-square and asymptotic stability of the numerical method applied to the test SDE (2.3). In particular, for the numerical mean-square stability, we have [Hig00]

$$\lim_{n\to\infty} \mathbb{E}(|X_n|^2) = 0 \iff (p, q) \in \mathcal{S}_{num}^{MS} \text{ where } \mathcal{S}_{num}^{MS} := \left\{(p, q) \in \mathbb{C}^2 \,; \mathbb{E}|R(p, q, \xi)|^2 < 1\right\}, \tag{2.7}$$

and for the numerical asymptotic stability, assuming $R(p, q, \xi) \neq 0$ with[1] probability 1 and $\mathbb{E}((\log |R(p, q, \xi)|)^2) < \infty$, it is shown in [Hig00, Lemma 5.1] the equivalence

$$\lim_{n\to\infty} |X_n| = 0 \text{ with probability 1} \iff (p, q) \in \mathcal{S}_{num}^{AS}, \tag{2.8}$$

with $\mathcal{S}_{num}^{AS} := \left\{(p, q) \in \mathbb{C}^2 \ ; \ \mathbb{E}(\log |R(p, q, \xi)|) < 0\right\}$.

Mean-square A-stability and L-stability We denote $\mathcal{S}_{num}^{AS}, \mathcal{S}_{num}^{MS}$, respectively, the above domains of asymptotic and mean-square stability. A numerical integrator is called

- mean-square A-stable if $S_{\text{SDE}}^{MS} \subseteq \mathcal{S}_{num}^{MS}$;

[1] Note that if $R(p, q, \xi) = 0$ with a non-zero probability, then (2.6) is numerically asymptotically stable.

- mean-square L-stable, if it is mean-square A-stable and if $\mathbb{E}(|R(p_k, q_k, \xi)|^2) \to 0$ holds for all sequences $(p_k, q_k) \in S^{MS}_{SDE}$ with $\Re(p_k) \to -\infty$.

If we restrict $(p, q) \in \mathbb{R}^2$ then the domains of mean-square or asymptotic stability are called regions of stability. Applied to the scalar test equation (2.3), the Milstein-Talay method

Milstein-Talay method, see (1.29)

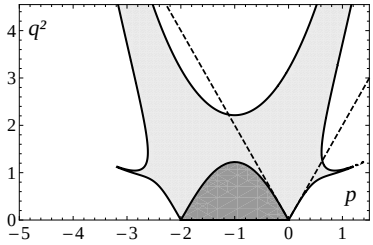


Figure 2.1: Mean-square stability region (dark gray) and asymptotic stability region (dark and light grays) of the explicit second order Milstein-Talay method with stability function (2.9).

(1.29) (and also its derivative-free weak formulation (2.14) introduced in the next Section) has a stability function (2.6) satisfying

$$\mathbb{E}(|R(p, q, \xi)|^2) = \left|1 + p + \frac{p^2}{2}\right|^2 + |1 + p|^2|q|^2 + \frac{|q|^4}{2}. \tag{2.9}$$

It can be seen in Figure 2.1 that this scheme has restricted mean-square and asymptotic stability regions in the (p, q)-plan (here with real values). This is expected for classical explicit methods and our goal is to introduce new weak second order integrators with extended stability domains to avoid severe timestep size restrictions for stiff SDEs. Here, the dotted lines in Figure 2.1 indicate the boundaries of the mean-square stability domain (2.4) and asymptotic stability domain (2.5) for the exact solution of (2.3).

2.2 Efficient derivative free explicit Milstein-Talay method

We briefly discuss the weak order two Milstein-Talay method (1.29) and explain an efficient implementation of (1.29) that will be helpful to understand our new stabilized stochastic integrators. First it is well-known that one can replace the stochastic integrals $I_{r,0}, I_{0,r}, I_{q,r}$ in (1.29) by discrete random increments without altering the weak order two. Consider independent discrete random variables $\chi_r, \xi_r,\ r = 1, \ldots, m$ satisfying

$$\mathbb{P}(\chi_r = \pm 1) = 1/2, \quad \mathbb{P}(\xi_r = \pm\sqrt{3}) = 1/6, \quad \mathbb{P}(\xi_r = 0) = 2/3, \tag{2.10}$$

then both $I_{r,0}$ and $I_{0,r}$ can be replaced by $h^{3/2}\xi_r$ and $I_{q,r}$ can be replaced by

$$J_{q,r} = \begin{cases} h(\xi_r\xi_r - 1)/2 & \text{if } q = r \\ h(\xi_q\xi_r - \chi_q)/2 & \text{if } r < q \\ h(\xi_q\xi_r + \chi_r)/2 & \text{if } r > q. \end{cases} \tag{2.11}$$

The weak approximation (2.11) involving $2m - 1$ discrete random variables was first proposed in [Mil86] (see also [MT04, p. 96, eq. (1.25)]). The weak second order method (1.29)

with discrete random increments then reads (see e.g. [MT04, p. 103, eq. (2.18)])

$$\begin{aligned}\hat{X}_1 &= X_0 + hf(X_0) + \sqrt{h}\sum_{r=1}^m g^r(X_0)\xi_r + \sum_{q,r=1}^m (g^r)'(X_0)g^q(X_0)J_{q,r} \\ &+ \frac{h^2}{2}\Big(f'(X_0)f(X_0) + \frac{1}{2}\sum_{r=1}^m f''(X_0)(g^r(X_0), g^r(X_0))\Big) \qquad (2.12)\\ &+ \sum_{r=1}^m \Big((g^r)'(X_0)f(X_0) + \frac{1}{2}\sum_{q=1}^m (g^r)''(X_0)(g^q(X_0), g^q(X_0)) + f'(X_0)g^r(X_0)\Big)\frac{h^{3/2}\xi_r}{2}.\end{aligned}$$

Using additional Runge-Kutta stages allows to remove $f'f$, $f'g^r$, $f''(g^r, g^r)$ without altering the weak order two of (2.12). Next, we use the following approximation first proposed in [Rößß09] to construct efficient derivative free second order methods,

$$\begin{aligned}\sum_{q,r=1}^m (g^r(X_0))'g^q(X_0)J_{q,r} &= \frac{1}{2}\sum_{r=1}^m \Big[g^r\Big(X_0 + \sum_{q=1}^m g^q(X_0)J_{q,r}\Big) - g^r\Big(X_0 - \sum_{q=1}^m g^q(X_0)J_{q,r}\Big)\Big] \\ &+ \mathcal{O}(h^3). \qquad (2.13)\end{aligned}$$

Again, this approximation does not alter the weak second order of the method and requires only 3 evaluations of each function g^r. In contrast, a naive finite difference approximation e.g., $\frac{1}{2h}\sum_{q,r=1}^m \Big[g^r\Big(x + hg^q(x)\Big) - g^r\Big(x - hg^q(x)\Big)\Big] J_{q,r}$, would require $2m+1$ evaluations of each function g^r at the points $x, x \pm hg^q(x)$. Finally, we naturally arrive to the following scheme for general systems of Itô SDEs (1.22).

Algorithm 2.2.1 (Derivative-free Milstein-Talay integrator of weak order 2)
Given X_0, compute X_1 explicitly as follows.

$$\begin{aligned}K_1 &= X_0 + hf(X_0), \quad K_2 = K_1 + \sqrt{h}\sum_{r=1}^m g^r(X_0)\xi_r, \\ X_1 &= X_0 + \frac{h}{2}\Big(f(X_0) + f(K_2)\Big) \\ &+ \frac{1}{2}\sum_{r=1}^m \Big(g^r\Big(X_0 + \sum_{q=1}^m g^q(X_0)J_{q,r}\Big) - g^r\Big(X_0 - \sum_{q=1}^m g^q(X_0)J_{q,r}\Big)\Big) \qquad (2.14)\\ &+ \frac{\sqrt{h}}{2}\sum_{r=1}^m \Big(g^r\big(\frac{X_0+K_1}{2} + \sqrt{\frac{h}{2}}\sum_{q=1}^m g^q(X_0)\chi_q\big) + g^r\big(\frac{X_0+K_1}{2} - \sqrt{\frac{h}{2}}\sum_{q=1}^m g^q(X_0)\chi_q\big)\Big)\xi_r.\end{aligned}$$

Each step of the above scheme necessitates only five evaluations of the diffusion functions g^r, $r = 1, \dots, m$, independently of the dimension m. The method (2.14) – a modification of the second order method in [KP92, eq. (2.7) Chap. 14] – seems not to have appeared in the literature, in particular the finite difference discretisation in the last line of (2.14) seems new. The following proposition states that its weak order of accuracy is two. The idea of the proof is to study the difference after one step of weak order two scheme (2.12) and the modified version (2.14). It relies on standard arguments using Remark 1.2.1. It will be useful for the construction of high weak order integrators for stiff SDEs.

Theorem 2.2.2 *Consider the system of SDEs* (1.22) *with $f, g^r \in C_P^6(\mathbb{R}^N, \mathbb{R}^N)$, Lipschitz continuous. Then the derivative free Milstein-Talay method* (2.14) *for the approximation of* (1.22) *satisfies*

$$|\mathbb{E}(\phi(X(nh))) - \mathbb{E}(\phi(X_n))| \le Ch^2, \quad 0 \le nh \le T$$

for all $\phi \in C_P^6(\mathbb{R}^N, \mathbb{R})$, where C is independent of n, h.

2.3 Diagonally implicit integrators for SDEs

This section summarizes the work [AVZ13b]. Based on the derivative-free Milstein-Talay integrator (2.14), we introduce the following integrator of weak order two for the integration of (1.22). We highlight that the integrator is drift-implicit, which means that it is implicit with respect to the drift function f, but explicit with respect to the diffusion functions $g^r, r = 1, \dots, m$. Recall that such explicitness with respect to the diffusion functions is a desirable property to avoid stability issues (see Remark 1.2.8).

Algorithm 2.3.1 *(S-SDIRK: diagonally implicit Runge-Kutta method of weak second order) Given X_0, compute X_1 as follows.*

$$
\begin{aligned}
K_1 &= X_0 + \gamma h f(K_1),\\
K_2 &= X_0 + (1-2\gamma) h f(K_1) + \gamma h f(K_2),\\
K_1^* &= X_0 + \beta_1 \gamma h f(K_1) + \beta_2 \gamma h f(K_2),\\
K_2^* &= X_0 + \gamma h f(K_1) + D^{-1}(K_1^* - X_0),\\
K_3^* &= K_1^* + \beta_3 h f(K_2^*),\\
X_1 &= X_0 + \frac{h}{2} f(K_1) + \frac{h}{2} f\Big(K_2 + \sqrt{h}\sum_{r=1}^m g^r(K_2^*)\xi_r\Big)\\
&+ \frac{1}{2}\sum_{r=1}^m \Big[g^r\Big(K_2^* + \sum_{q=1}^m g^q(K_2^*) J_{q,r}\Big) - g^r\Big(K_2^* - \sum_{q=1}^m g^q(K_2^*) J_{q,r}\Big)\Big]\\
&+ \frac{\sqrt{h}}{2}\sum_{r=1}^m \Big[g^r\Big(K_3^* + \sqrt{\frac{h}{2}}\sum_{q=1}^m g(K_2^*)\chi_q\Big) + g^r\Big(K_3^* - \sqrt{\frac{h}{2}}\sum_{q=1}^m g(K_2^*)\chi_q\Big)\Big]\xi_r
\end{aligned}
\tag{2.15}
$$

where $\beta_1 = \frac{2-5\gamma}{1-2\gamma}$, $\beta_2 = \frac{\gamma}{1-2\gamma}$, $\beta_3 = \frac{1}{2} - 2\gamma$, and $\xi_r, \chi_r, J_{q,r}$ satisfy (2.10) *and* (2.11) *respectively.*

For $\gamma = 0$, we have $K_1^* = K_2^* = X_0$ and $K_3^* = X_0 + (h/2) f(X_0)$ and we recover the explicit Milstein-Talay method (2.14). For the stage K_2^*, we use D^{-1} to stabilize $K_1^* - X_0$, where $D = I - \gamma h f'(X_0)$. This stabilization procedure is well-known in ODEs (to stabilize the error estimator of an integrator) and has been introduced by Shampine [HW96, Sect. IV.8], its use for SDEs is motivated in Remark 2.3.2 below. We emphasize that it does not represent a computational overhead as the LU-factorization of D needed to compute $D^{-1}(K_1^* - X_0)$ is already available from the solution of the nonlinear system for the stages (K_1, K_2) (see Remark 2.3.3).

We consider two choices for γ that yield mean-square A-stable integrators:

- the S-SDIRK(2,2) method for the value $\gamma = 1 - \frac{\sqrt{2}}{2}$ which gives a weak order 2 A-stable method with deterministic order 2;
- the S-SDIRK(2,3) method for the value $\gamma = \frac{1}{2} + \frac{\sqrt{3}}{6}$ which gives a weak order 2 A-stable method with deterministic order 3.

The value of γ for the S-DIRK(2,3) yields in the deterministic case a method of order 3 which is strongly A-stable, i.e. $|R(\infty)| < 1$, while the value of γ for the S-DIRK(2,2) yields a method of order 2 which is L-stable, i.e. it is A-stable and $R(\infty) = 0$. L-stability is desirable in the case of very stiff deterministic problems as it permits to damp the very high frequencies.

Remark 2.3.2 *We observe that by removing the term involving D^{-1} in the S-SDIRK methods* (2.15)*, the denominators of the stability functions of the internal stages K_j^* would scale at best as $(1-\gamma p)^2$. The resulting methods would no longer be mean-square A-stable.*

Standard arguments permit to show that the integrator (2.15) has second weak order of accuracy for general systems of SDEs: it satisfies the statement of Theorem 2.2.2.

Complexity In addition to the solution of the deterministic two stage SDIRK method (which yields the stages (K_1, K_2)) one step of the scheme (2.15) costs one evaluation of the drift function f, 5 evaluations of each diffusion functions g^r, and the generation of $2m$ random variables. The cost is similar to the diagonally implicit methods proposed in [DR09a] (in particular the number of evaluation of the diffusion functions g^r, $r = 1, \dots, m$ is independent of the number of Wiener processes m).

Remark 2.3.3 *We emphasize that the computation of $D^{-1}(K_1^* - X_0)$ in the scheme* (2.15) *does not represent any computational overhead. Indeed, as for any deterministic or stochastic diagonally implicit method [HW96, DR09a], the usual procedure for evaluating K_1, K_2 is to compute the LU-factorization of $D = I - \gamma h f'(X_0)$ ($f'(X_0)$ is usually further approximated by finite differences) and make the quasi-Newton iterations*

$$LU(K_i^{k+1} - K_i^k) = -K_i^k + X_0 + \delta_{2i}(1-2\gamma)hf(K_1) + \gamma h f(K_i^k),\ i = 1, 2, \tag{2.16}$$

where δ_{2i} is the Kronecker delta function. The same LU-factorization is then used to compute $D^{-1}(K_1^* - X_0)$ by solving

$$LUY = K_1^* - X_0,$$

whose cost in negligible: the cost of evaluating K_2^* together with K_3^* is the same as one iteration of (2.16)*.*

The following Theorem states that S-SDIRK(2,2) is not only a mean-square A-stable integrator, but also a mean-square L-stable integrator, as defined in Section 2.1.2.

Theorem 2.3.4 *The integrator S-SDIRK(2,2) is mean-square L-stable.*

The main idea of the proof of Theorem 2.3.4 is to note that $\mathbb{E}(|R(p, q, \xi)|^2)$ is an increasing function of $|q|^2$, thus the mean-square A-stability of the method is equivalent to

$$\sup_{\Re z < 0} \mathbb{E}(|R(z, \sqrt{-2\Re z}, \xi)|^2) \leq 1.$$

and for the mean-square L-stability one has to prove in addition

$$\sup_{\Re z < 0} \mathbb{E}(|R(z, \sqrt{-2\Re z}, \xi)|^2) \to 0 \text{ for } \Re z \to -\infty.$$

The above quantity can be studied as a function of $z = x + iy$. It turns out that its maximum with respect to y is achieved simply for $y = 0$.

Remark 2.3.5 *It can be checked numerically that the integrator S-SDIRK(2,3) is mean-square A-stable. A rigorous proof is however more tedious to derive because the deterministic stability function of the method does not decay to zero (note that the scheme is not L-stable).*

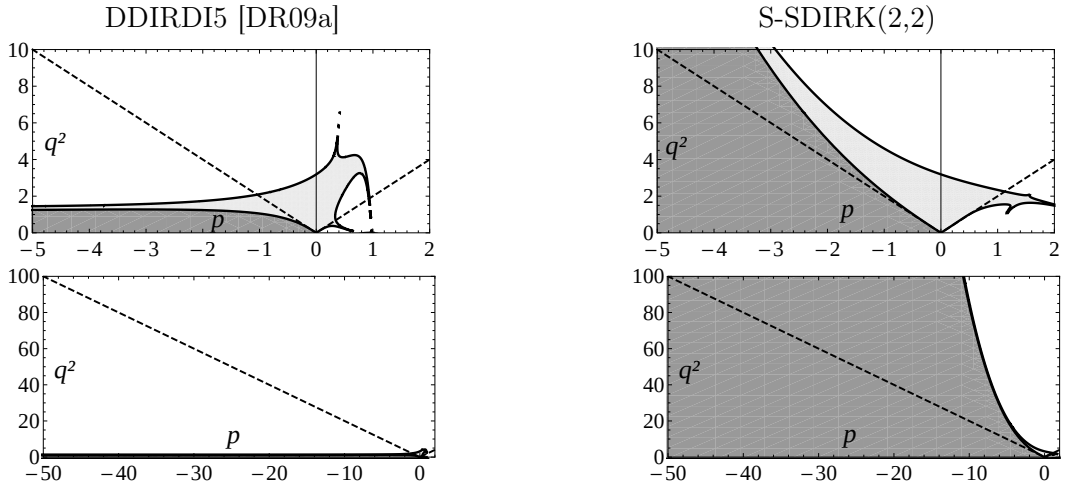


Figure 2.2: Mean-square stability region (dark gray) and asymptotic stability region (dark and light grays) for DDIRDI5 [DR09a] (left pictures) and S-SDIRK(2,2) (right pictures).

SDE test problem: $dX(t) = \lambda X(t)dt + \mu X(t)dW(t)$				
	mean-square	stepsize restriction for mean-square stability		
method	A-stability	$-\lambda = \mu^2 = 5$	$-\lambda = \mu^2 = 50$	$-\lambda = \mu^2 = 500$
Milstein-Talay (2.14)	no	$h \leq 0.236$	$h \leq 0.0236$	$h \leq 0.00236$
DDIRDI5 [DR09a]	no	$h \leq 0.246$	$h \leq 0.0246$	$h \leq 0.00246$
S-SDIRK (2,2) or (2,3)	yes	no restriction	no restriction	no restriction

Table 2.1: Comparison of mean-square stability constraints.

The mean-square A-stability of S-SDIRK(2,2) is illustrated in in Figure 2.2 (right pictures). We now exhibit the advantage of our method over the Milstein-Talay method (2.14) and the weak second order drift-implicit method DDIRDI5 considered in [DR09a] and suitable only for small noise regimes (left pictures in Fig. 2.2). In particular, we consider the linear test problem (2.3) and compare the behaviour of the three different methods for a range of parameters λ, μ for which the solution of (2.3) is mean-square stable. As we can see in Table 2.1, even for a moderate stiff problem ($-\lambda = \mu^2 = 5$) in contrast to the S-SDIRK methods introduced here, there is quite a severe stepsize restriction in order for the numerical solution to be mean-square stable for the Milstein-Talay and the DDIRDI5 methods. Furthermore, as expected we observe that the stepsize restriction for the other two methods becomes more severe as we increase the stiffness of the problem.

We finally compare the performance of the introduced stochastic integrators on a nonlinear stiff system of SDEs with a one-dimensional noise ($d = 2, m = 1$),

$$\begin{aligned} dX(t) &= \big(\alpha(Y(t)-1) - \lambda_1 X(t)(1-X(t))\big)dt - \mu_1 X(t)(1-X(t))dW(t), \\ dY(t) &= -\lambda_2 Y(t)(1-Y(t))dt - \mu_2 Y(t)(1-Y(t))dW(t), \end{aligned} \tag{2.17}$$

which is inspired from a one-dimensional population dynamics model [Gar88, Chap. 6.2]. Note that if we linearise (2.17) around the stationary solution $(X, Y) = (1, 1)$, for $\alpha = 0$ we recover (twice) the linear test problem (2.3). We take the initial conditions $X(0) = Y(0) = 0.95$ close to this steady state and use the parameters $\lambda_2 = -4$, $\mu_2 = 1$, $\alpha = 1$.

We take for the deterministic part of the problem the stiff parameter $\lambda_1 = -500$ and we shall consider for the noise parameter μ_1 either the stiff value $\mu_1 = \sqrt{500}$ or the non-stiff

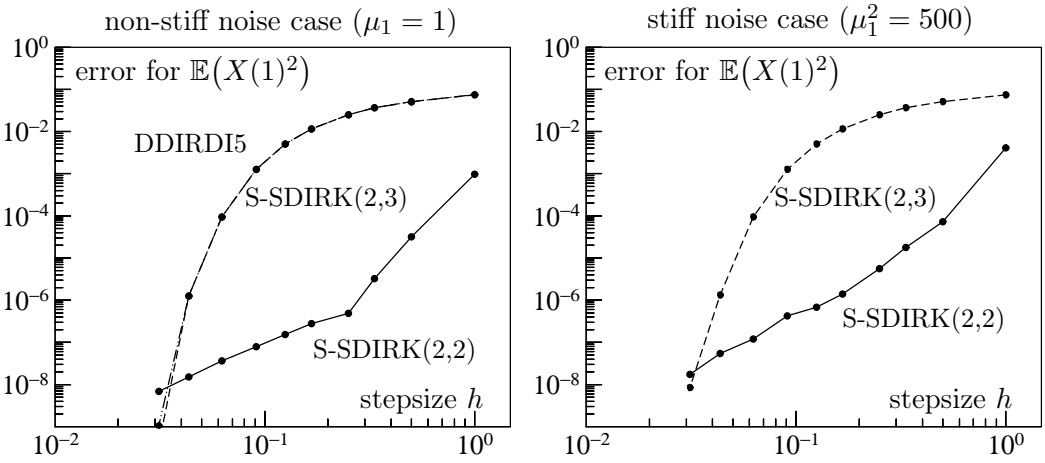


Figure 2.3: Weak convergence plots for the nonlinear stiff problem (2.17) for S-SDIRK(2,2) (solid line), S-SDIRK(2,3) (dashed line), DDIRDI5 (dashed-dotted-dotted line) [DR09a]. Error for $\mathbb{E}(X(1)^2)$ versus the stepsize h, where $1/h = 1, 2, 3, 4, 6, 8, 11, 16, 23, 32$. Averages over 10^8 samples.

value $\mu_1 = 1$. We plot in Figure 2.3 the errors for $\mathbb{E}(X(T)^2)$ at the final time $T = 1$ versus stepsizes h for the integrators DDIRDI5, S-SDIRK(2,2), S-SDIRK(2,3) taking the averages over 10^8 samples. Reference solutions where computed using the Milstein-Talay method (2.14) with stepsize $h = 10^{-4}$. We consider the two cases of a non-stiff noise ($\mu_1 = 1$) and a stiff noise ($\mu_1 = \sqrt{500}$). In the non-stiff noise case (left picture), the results of S-SDIRK(2,3) are nearly identical to those of DDIRDI5 with hardly distinguishable curves, while in the stiff noise case (right picture), the results for DDIRDI5 are not included because this method is unstable for the considered stepsizes, as predicted by the linear stability analysis (see the stepsize restrictions in Table 2.1). It is remarkable in both cases that S-SDIRK(2,2) is more than four magnitudes more accurate than S-SDIRK(2,3) for steps with size $\sim 10^{-1}$, a regime for which curves with slope two can be observed. We believe that the mean-square $L-$stability of the S-SDIRK(2,2) method is responsible for this behavior.

2.4 Explicit stabilized integrators for stiff stochastic problems

This section summarized the work [AVZ12]. The mean-square stability domain of an explicit stochastic integrator is always bounded because its stability function involves polynomial expressions. Such integrator is thus never mean-square A-stable. The aim of the Section is to present a new class of explicit stabilized integrators, whose mean-square stability domain size grows rapidly compared to the computational cost (quadratically with respect to the number of function evaluations of the method).

We first define for $a > 0$ the following "portion of the true mean-square stability region"

$$\mathcal{S}_a^{MS} = \{(p, q) \in (-a, 0) \times \mathbb{R} \ ; \ p + \tfrac{1}{2}|q|^2 < 0\}, \tag{2.18}$$

and define for a given method

$$\ell = \sup\{a > 0 \ ; \ \mathcal{S}_a^{MS} \subset \mathcal{S}_{num}^{MS}\}, \quad d = \sup\{a > 0 \ ; \ (-a, 0) \times \{0\} \subset \mathcal{S}_{num}^{MS}\}, \tag{2.19}$$

where d is the size of the stability domain along the deterministic p-axis (observe that $d \geq \ell$). For the Milstein-Talay methods (1.29) or (2.14), we have $\ell = 0, d = 2$. In contrast, the new S-ROCK2 methods introduced in the present section have values ℓ, d increasing quadratically with the stage parameter s. In turn, the ratio of stability versus work increases linearly, while for classical explicit methods, it is bounded.

Weak order one S-ROCK methods [AL08] For deterministic systems (1.1) of ODEs, a well-know stabilization procedure for the Euler method has been proposed in [VS80]. Its construction is based on the classical Chebyshev polynomials $T_s(\cos x) = \cos(sx)$. Given an integer $s \geq 1$, the number of stages, and a damping parameter $\eta \geq 0$, we define the following Runge-Kutta method (first order Chebyshev method) with step size h by the following explicit recursion

$$\begin{aligned} K_0 &= X_0, \quad K_1 = X_0 + h\frac{\omega_1}{\omega_0} f(K_0), \\ K_j &= 2h\frac{T_{j-1}(\omega_0)}{T_j(\omega_0)} f(K_{j-1}) + 2\omega_0 \frac{T_{j-1}(\omega_0)}{T_j(\omega_0)} K_{j-1} - \frac{T_{j-2}(\omega_0)}{T_j(\omega_0)} K_{j-2}, \quad j = 2, \ldots, s \\ X_1 &= K_s, \end{aligned} \tag{2.20}$$

where $\omega_0 = 1 + \frac{\eta}{s^2}$, $\omega_1 = \frac{T_s(\omega_0)}{T_s'(\omega_0)}$. Applied to the linear test problem $dX(t)/dt = \lambda X(t)$ the method (2.20) gives $X_1 = R_s(p)X_0$, where $p = \lambda h$ and where $R_s(p)$, called the stability function (polynomial) of the method, is given by $R_s(p) = T_s(\omega_0 + \omega_1 p)/T_s(\omega_0)$. We emphasize that (2.20) denotes in fact a family of methods indexed by the stage number s. A crucial property of the methods (2.20) is

$$|R_s(p)| \leq 1 \quad \text{for all } p \in (-d_s, 0), \tag{2.21}$$

with $d_s \simeq C \cdot s^2$, for s large enough, where C depends on the damping parameter η (for $\eta = 0$, $C = 2$). Thus, the length d_s of the stability domain

$$\mathcal{S} := \{p \in \mathbb{C}; |R(z)| \leq 1\} \tag{2.22}$$

of the methods increases quadratically with s on the negative real axis. This quadratic growth of the stability domain is the key feature of such methods compared to standard explicit integrators.

The idea for stabilizing the Euler-Maruyama (1.28) is now simply to damp its stability function $R(p, q, \xi) = 1 + p + q\xi$, obtained by applying (1.28) to (2.3) using $R_s(p)$ (with a value of the damping η optimized for each s, see [AL08]). The corresponding Runge-Kutta type scheme reads [AL08]

$$\begin{aligned} K_0 &= X_0, \quad K_1 = X_0 + h\frac{\omega_1}{\omega_0} f(K_0), \\ K_j &= 2h\frac{T_{j-1}(\omega_0)}{T_j(\omega_0)} f(K_{j-1}) + 2\omega_0 \frac{T_{j-1}(\omega_0)}{T_j(\omega_0)} K_{j-1} - \frac{T_{j-2}(\omega_0)}{T_j(\omega_0)} K_{j-2}, \quad j = 2, \ldots, s \\ X_1 &= K_s + \sum_{r=1}^{m} g^r(K_s)\Delta W_r. \end{aligned} \tag{2.23}$$

The method (2.23) is denoted S-ROCK(1/2,1) and has strong order 1/2 and weak order 1 for general systems of SDEs (1.22). Another method of strong and weak orders 1 has been considered in [AL08] in a one-dimensional context. Using the approximation (2.13)

from [Rößß09], a multi-dimensional derivative free version, denoted S-ROCK(1,1), can be obtained straightforwardly by replacing the last line in (2.23) by

$$X_1 = K_s + \sum_{r=1}^{m} g^r(K_s)\Delta W_r + \frac{1}{2}\sum_{r=1}^{m}\Big(g^r\big(K_s + \sum_{q=1}^{m} g^q(K_s)I_{q,r}\big) - g^r\big(K_s - \sum_{q=1}^{m} g^q(K_s)I_{q,r}\big)\Big),$$

where $I_{q,r}$ are defined in (1.30) and by considering a larger damping η as discussed in [AL08] (see also the related work [KB13]). It turns out that S-ROCK(1/2,1) and S-ROCK(1,1) include a portion of the true mean-square stability region that scales like $\ell_s \simeq 0.33 \cdot s^2$ and $\ell_s \simeq 0.19 \cdot s^2$, respectively.

Second order stabilization Similarly to the weak order one S-ROCK methods, the idea to design new integrators is to stabilize a weak second order method, and we shall consider the non-stiff integrator (2.14). We start with a deterministic stabilized second order Chebyshev method. Recall that the derivation of optimal stability functions suitable for the stabilization of second order (deterministic methods) is a non trivial task and various strategies have been proposed [Leb89, VS80, AM01, Abd02]. We choose here the second order orthogonal Runge-Kutta Chebyshev methods (ROCK2) introduced in [AM01]. The idea is to search for a stability polynomial $R_s(p) = w_2(p)P_{s-2}(p)$, where $P_{s-2}(p)$ is a member family of polynomials $\{P_j(z)\}_{j\geq 0}$ orthogonal with respect to the weight function $\frac{w_2(x)^2}{\sqrt{1-x^2}}$. The polynomial P_{s-2} has degree $s-2$, while w_2 is a positive polynomial of degree two (depending on s). One constructs the polynomials w_2 such that R_s satisfies [AM01]

$$R_s(p) = 1 + p + \frac{p^2}{2} + \mathcal{O}(p^3), \tag{2.24}$$

together with a large stability interval along the negative real axis (2.21), increasing as $d_s \simeq 0.81 \cdot s^2$. Thanks to the recurrence relation of the orthogonal polynomials $\{P_j(z)\}_{j\geq 0}$, a method of order two for (1.1) based on a recurrence formula can be constructed [2]

$$\begin{aligned}
K_0 &= X_0, \quad K_1 = K_0 + \mu_1 hf(K_0),\\
K_j &= \mu_j hf(K_{j-1}) - \nu_j K_{j-1} - \kappa_j K_{j-2}, \quad j = 2, \cdots, s-2,\\
K_{s-1} &= K_{s-2} + 2\tau hf(K_{s-2}),\\
X_1 &= K_{s-2} + \left(2\sigma - \frac{1}{2}\right) hf(K_{s-2}) + \frac{1}{2} hf(K_{s-1}).
\end{aligned} \tag{2.25}$$

The parameters μ_j, k_j (depending on s) are obtained from the three-term recurrence relation [AM01, eq. (24)-(25)] of the orthogonal polynomials $\{P_j(z)\}_{j\geq 0}$, while σ, τ (that also depend on s) satisfy $w_2(p) = 1 + 2\sigma p + \tau p^2$ and are chosen such that (2.24) holds.

In preparation for the extension of the ROCK2 methods to stochastic problems, we explain a novel strategy to introduce damping in the scheme (2.25). The idea is to consider the following scheme for a fixed scalar parameter α.

$$\begin{aligned}
K_0 &= X_0, \quad K_1 = K_0 + \alpha\mu_1 hf(K_0),\\
K_j &= \alpha\mu_j hf(K_{j-1}) - \nu_j K_{j-1} - \kappa_j K_{j-2}, \quad j = 2, \cdots, s-2,\\
K_{s-1} &= K_{s-2} + 2\tau_\alpha hf(K_{s-2})\\
X_1 &= K_{s-2} + \left(2\sigma_\alpha - \frac{1}{2}\right) hf(K_{s-2}) + \frac{1}{2} hf(K_{s-1}).
\end{aligned} \tag{2.26}$$

[2] The two last stages of the method are written is a slightly different way as in the ROCK2 method [AM01, Equ. (26-27)] as (2.25) is more convenient for an extension to stochastic integrators. We emphasize that it has the same order and similar stability properties as the ROCK2 method [AM01, Equ. (26-27)].

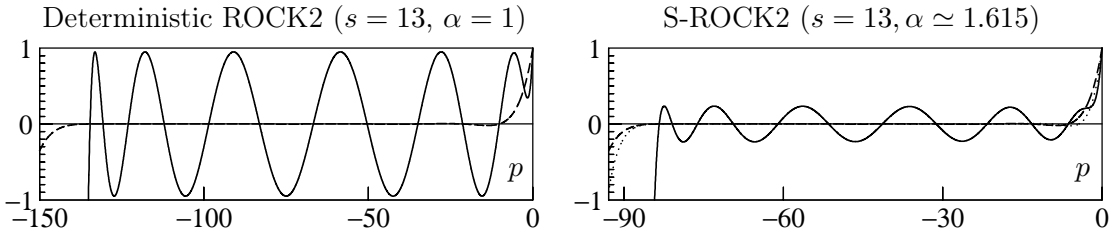


Figure 2.4: Comparison of polynomials involved in ROCK2 and S-ROCK2 for $s = 13$. Polynomials $R_{s,\alpha}$ (solid lines), $P_{s-2}(\alpha p)$ (dashed lines). We also include the polynomial $P_s(\alpha p)$ in the right picture (dotted lines).

Note that for $\alpha = 1$, we recover the original ROCK2 method (2.25). Applied to the linear test problem $dX/dt = \lambda X, X(0) = X_0$ this method yields (setting $p = h\lambda$ and $X_0 = 1$)

$$X_1 = (1 + 2\sigma_\alpha p + \tau_\alpha p^2) P_{s-2}(\alpha p) =: R_{s,\alpha}(p). \tag{2.27}$$

Lemma 2.4.1 *The method* (2.26) *has second order for the system of ODEs* (1.1) *for any* α *provided*

$$\sigma_\alpha = \frac{1-\alpha}{2} + \alpha\sigma, \quad \tau_\alpha = \frac{(\alpha-1)^2}{2} + 2\alpha(1-\alpha)\sigma + \alpha^2\tau. \tag{2.28}$$

In Figure 2.4, we plot, for $s = 13$, the polynomials $P_{s-2}(\alpha p)$ and $R_{s,\alpha}(p)$ (defined in (2.27)) involved in the standard ROCK2 method ($\alpha = 1$, left picture) and the S-ROCK2 method ($\alpha \simeq 1.615$ right picture) introduced in the next section. It can be seen that increasing α reduces the amplitude of the oscillations of $R_{s,\alpha}(p)$. The appropriate choice of α is discussed below.

We are now in position to introduce our new explicit stabilized integrator, obtained by stabilizing the stochastic Milstein-Talay method (2.14) with the modified deterministic ROCK2 method (2.26).

Algorithm 2.4.2 (S-ROCK2 integrator of weak order two) *Given* X_0*, compute* X_1 *as follows.*

$$\begin{aligned}
K_0 &= X_0, \quad K_1 = K_0 + \alpha\mu_1 h f(K_0), \\
K_j &= \mu_j \alpha h f(K_{j-1}) - \nu_j K_{j-1} - \kappa_j K_{j-2}, \quad j = 2, \cdots, s, \\
K^*_{s-1} &= K_{s-2} + 2\tau_\alpha h f(K_{s-2}) + \sqrt{h}\sum_{r=1}^{m} g^r(K_s)\xi_r, \\
X_1 &= K_{s-2} + \Big(2\sigma_\alpha - \frac{1}{2}\Big) h f(K_{s-2}) + \frac{1}{2} h f(K^*_{s-1}) \\
&+ \frac{1}{2}\sum_{r=1}^{m}\Big(g^r\Big(K_s + \sum_{q=1}^{m} g^q(K_s) J_{q,r}\Big) - g^r\Big(K_s - \sum_{q=1}^{m} g^q(K_s) J_{q,r}\Big)\Big) \\
&+ \frac{\sqrt{h}}{2}\sum_{r=1}^{m}\Big(g^r\big(K_{s-1} + \sqrt{\frac{h}{2}}\sum_{q=1}^{m} g^q(K_s)\chi_q\big) + g^r\big(K_{s-1} - \sqrt{\frac{h}{2}}\sum_{q=1}^{m} g^q(K_s)\chi_q\big)\Big)\xi_r.
\end{aligned} \tag{2.29}$$

where $\alpha = 1/(2P'_{s-1}(0))$ *and* $\sigma_\alpha, \tau_\alpha$ *are given by* (2.28)*. Here, the constants* $\mu_j, \nu_j, \kappa_j, \sigma, \tau$ *depend on* s *and are the same as for the standard deterministic ROCK2 integrator* (2.25)*.*

integrator	work			stability	
	$\#f$	$\#g^r$	#random	d_s	ℓ_s
s steps of Milstein-Talay (2.14)	$2s$	$5s$	$2ms$	$2s$	0
one step of S-ROCK2 (2.29)	$s+2$	5	$2m$	$\simeq 0.42(s+2)^2$	$\simeq 0.42(s+2)^2$

Table 2.2: Computational complexity for an SDE in dimensions N (drift) and m (diffusion).

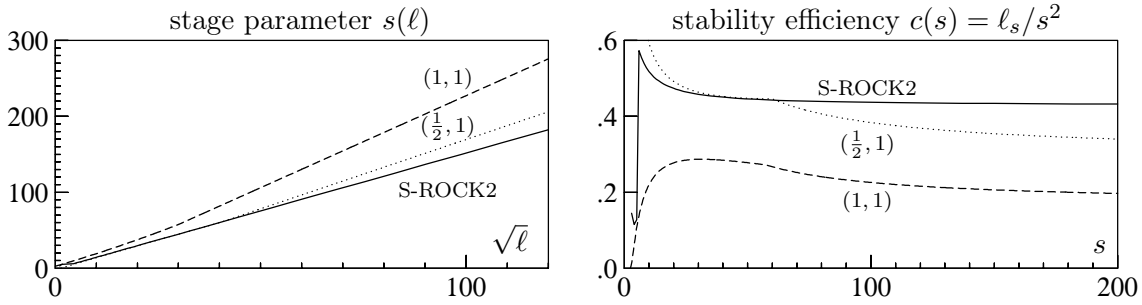


Figure 2.5: Comparison of S-ROCK2 (solid lines) and the weak order one S-ROCK methods $(1,1)$ (dashed lines), $(\frac{1}{2},1)$ (dotted lines). Left picture: optimal stage parameter s as a function of $\sqrt{\ell}$, where ℓ is given by (2.19). Right picture: stability efficiency $c(s) = \ell_s/s^2$.

Numerical computations show that the S-ROCK2 method includes a portion of the true mean-square stability region $\mathcal{S}_\ell^{MS}$ that grows with the stage number as $\ell_{\mathrm{S-ROCK2}} \simeq 0.42(s+2)^2$. The computational complexity of one step of the S-ROCK2 method with stepsize h is reported in Table 2.2 and compared to s steps with stepsize h/s of the weak second order Milstein-Talay method (2.14). The main feature of our S-ROCK2 integrators is that the mean-square stability region sizes ℓ_s, d_s grow quadratically with respect to the computational work $\#f + \#g^r$, while $\ell_s = 0$ and d_s grows only linearly for the standard explicit method.

In Figure 2.5 we plot the length ℓ defined in (2.19) of the portion of the true mean-square stability region $\mathcal{S}_\ell^{MS}$ as a function of the number of stages used. As we can see the behaviour of the S-ROCK2 method is $\ell \simeq Cs^2$ similarly to the S-ROCK methods of weak order 1 [AL08, AC08]. Furthermore, once can also see that the S-ROCK2 method is actually more efficient from a stability point of view, since the stability efficiency factor $c(s) = \ell_s/s^2$ converges numerically to about 0.42 for large s, which is larger than the S-ROCK(1/2,1/2) and S-ROCK(1,1) values of 0.33 and 0.19, respectively.

Standard arguments show that the S-ROCK2 method (2.29) has weak second order of accuracy: it satisfies the statement of Theorem 2.2.2 for general systems of SDEs.

Example: electric potential in a neuron Although our analysis applies only to systems of SDEs, we consider here an SPDE model for the propagation of an electric potential $V(x,t)$ in a neuron [Wal86]. This potential is governed by a system of non-linear PDEs called the Hodgkin-Huxley equations [HH52], but in certain ranges of values of V, this system of PDEs can be well approximated by the cable equation [Wal86]. In particular, if the neuron is subject to a uniform input current density over the dendrites and if certain geometric constraints are satisfied, then the electric potential satisfies the following linear cable equation with uniform input current density.

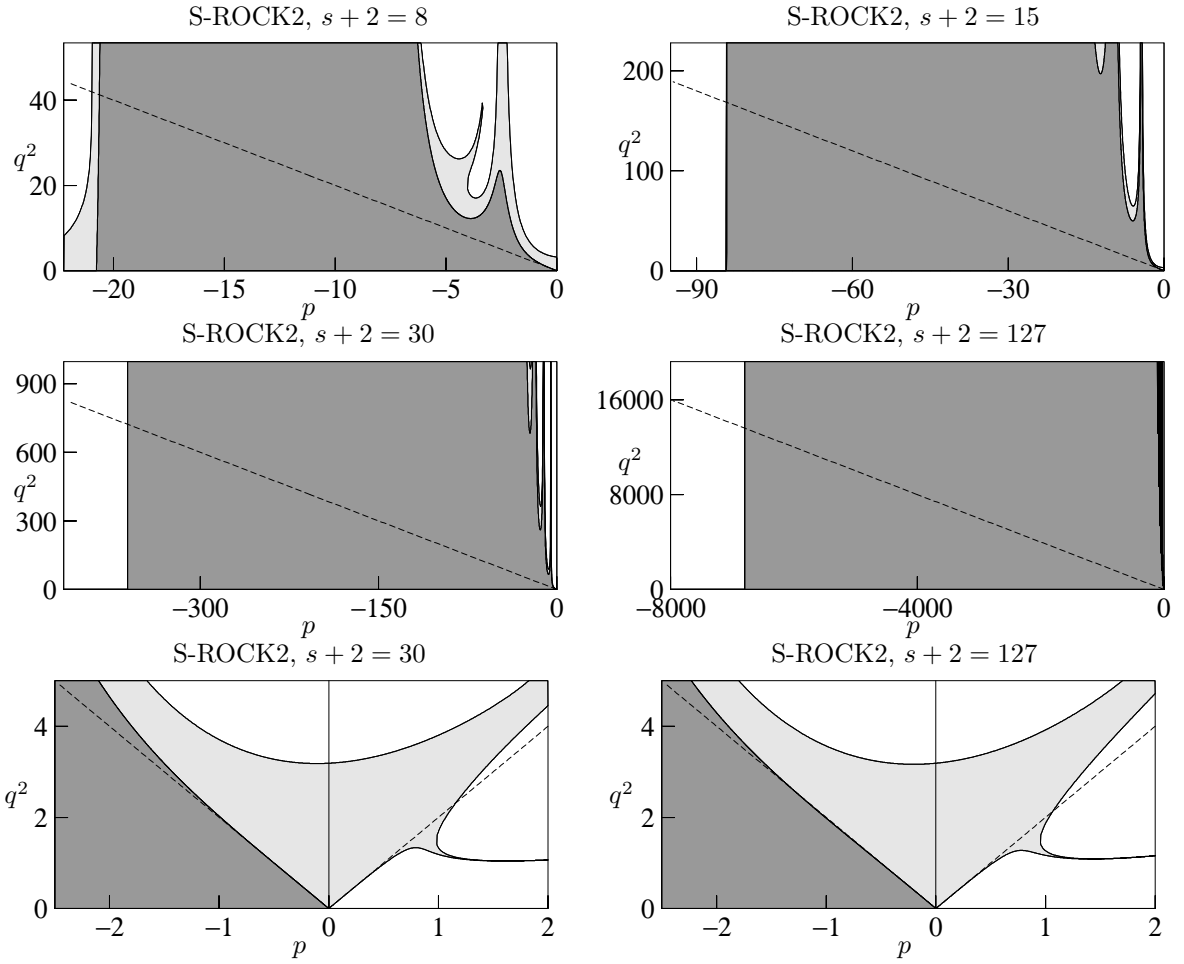


Figure 2.6: Mean-square stability regions (dark gray) and asymptotic stability regions (dark and light grays) of S-ROCK2 for $s+2=8,15,30$, and 127 stages, respectively.

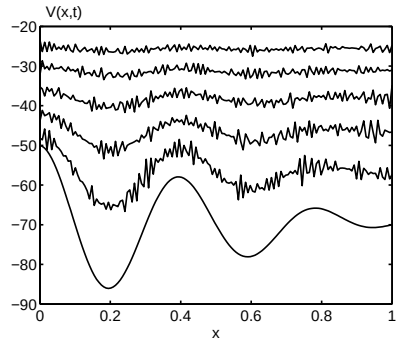
$$\begin{aligned}\frac{\partial V}{\partial t}(x,t) &= \nu\frac{\partial^2 V}{\partial x^2}(x,t) - \beta V(x,t) + \sigma(V(x,t)+V_0)\dot{W}(x,t), \quad 0\le x,t\le 1, \quad (2.30)\\ \frac{\partial V}{\partial x}(0,t) &= \frac{\partial V}{\partial x}(1,t) = 0, \quad t>0, \quad V(x,0)=V_0(x), \quad 0\le x\le 1,\end{aligned}$$

where $\dot{W}(x,t) = \frac{\partial^2}{\partial x \partial t} w(x,t)$ is a space-time white noise meant in the Itô sense. Here we have assumed that the distance between the origin (or soma) to the dendritic terminals is 1, and that the soma is located at $x=0$. Furthermore, the white noise term is describing the effect of the arrival of random impulses and the multiplicative noise structure depicts the fact that the response of the neuron to a current impulse may depend on a local potential [Wal86]. The quantity of interest is the threshold time $\tau = \inf\{t>0; V(t,0)>\lambda\}$, since when the potential at the soma (somatic depolarization) exceeds the threshold λ the neuron fires an action potential.

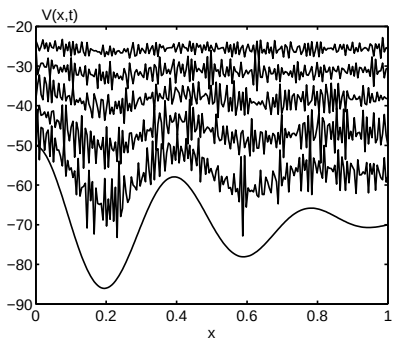
The SPDE (2.30) yields, after space discretization with finite differences [DG00] the following stiff system of SDEs where $V(x_i,t)\approx u_i$, with $x_i = i\Delta x$, $\Delta x = 1/N$,

$$du_i = \nu\frac{u_{i+1}-2u_i+u_{i-1}}{\Delta x^2}dt - \beta u_i dt + \sigma\frac{u_i+V_0}{\sqrt{\Delta x}}dw_i, \quad i=0,\dots,N, \qquad (2.31)$$

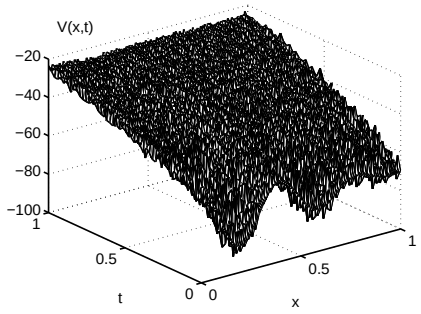
where the Neumann boundary condition requires $u_{-1}=u_1$ and $u_{N+1}=u_{N-1}$. Here $w_0,\dots w_N$ are independent standard Wiener processes, and dw_i indicates Itô noise. We



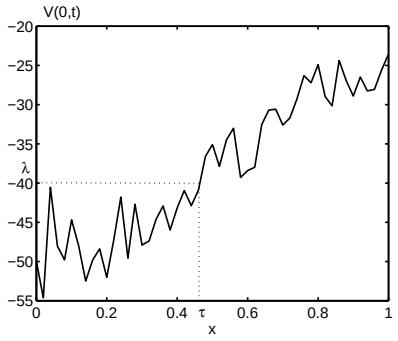
(a) $\Delta t = 1/50$, $\Delta x = 1/150$, fixed t.



(b) $\Delta t = 1/50$, $\Delta x = 1/200$, fixed t.



(c) $\Delta t = 1/50$, $\Delta x = 1/200$. Solution $V(x,t)$ as a function of x, t.



(d) $\Delta t = 1/50$, $\Delta x = 1/200$, fixed $x = 0$.

Figure 2.7: Samples of realisations of the SPDE (2.30) (discretized in space) using S-ROCK2 with $s + 2 = 8$ stages (resp. 11) for $\Delta x = 1/150$ (resp. $\Delta x = 1/200$). Figures (a),(b): solutions as functions of x at fixed times $t = 0, 0.2, 0.4, \ldots, 1.0$ (increasing with time, from bottom to top). Figure (d): solution as a function of t for $x = 0$.

consider the initial condition $V_0(x) = -70 + 20\cos(5\pi x)(1-x)$ and the constants $\nu = 10^{-2}$, $\sigma = 4 \cdot 10^{-3}$, $\beta = 1$, $V_0 = 10$, $\lambda = -40$. We consider the time interval $(0, T)$ with $T = 1$. Note that the noise in (2.31) is in diagonal form which permits to simplify the formulation of the method (see [AVZ12, Rem. 3.3]).

2.5 A "swiss-knife" integrator for stiff (stochastic) diffusion-advection-reaction problems

This section summarizes the work in [AV13b]. We introduce a new partitioned implicit-explicit integrator, called PIROCK, based on the explicit second order orthogonal Runge-Kutta Chebyshev method (ROCK2) introduced in [AM01] and combining ideas from [VSH04, AL08, AVZ12, AVZ13b, Zbi11]. We derive a single algorithm that can combine a diffusion term F_D with any combination of advection and/or reaction terms F_A, F_R and that can also treat Itô stochastic systems of the form

$$\dot{y} = F(y) = F_D(y) + F_A(y) + F_R(y) + \sum_{j=1}^{m} F_G^j(y)\dot{\xi}_j, \qquad y(0) = y_0, \tag{2.32}$$

where $\xi_j, j = 1, \ldots, m$ are independent one-dimensional Wiener processes. The main idea of the new method is to modify the finishing procedure of the standard ROCK2 method

[AM01], i.e. the final stages of ROCK2 used to achieve the order two of accuracy. We introduce a partitioned RK method, where the diffusion terms F_D and advection terms F_A are treated explicitly, while the reaction terms F_R are treated implicitly.

Compared to similar existing stabilized methods, the PIROCK method has the following features:

- for problems with stiff reactions, the number of function evaluations of the reaction terms F_R (solved implicitly) is independent of the stage number s used to handle the stiffness of the diffusion terms F_D (in contrast, the number of implicit stages in each step of the IRKC method is equal to s);

- for advection dominated problem,[3] the PIROCK method is more efficient than the RKC or ROCK2 solvers as it has better stability in the imaginary direction and requires a number of evaluations of the advection terms that is independent of the stage number of the method; compared to the PRKC method [Zbi11], the PIROCK method has larger stability domains on both the real and the imaginary parts;

- for problems with expensive evaluation of (non-stiff) reaction terms PIROCK is more efficient than RKC [VS80] or ROCK2 [AM01] as the number of evaluation of the reaction terms is independent of the stage number of the method; for such problems, it is comparable to the PRKC method [Zbi11] but has larger stability along the negative real axis;

- for problems involving white noise, it is more efficient than previously constructed S-ROCK methods [AC08, AL08], as PIROCK has a larger mean-square stability domain;

- it is the first explicit stabilized integrator that can treat non-symmetric diffusion operators: in the case of a non-symmetric differential operator the eigenvalues of the Jacobian of F_D are typically located in a sector

$$S_\theta = \{-\rho e^{i\tau}; \rho \geq 0, -\theta \leq \tau \leq \theta\} \tag{2.33}$$

  of the left half complex plane, where $\theta \leq \pi/2$ is the angle of this sector. The PIROCK integrator can deal with for large angles up to $\theta = \pi/4$, whereas standard stabilized integrators like RKC and ROCK2 can be applied only if θ is very small.

The proposed PIROCK algorithm is versatile and efficient (hence the " swiss-knife") in handling problems such as (1.1) for various regimes with a single code. It is fully adaptive and requires no tuning from the user. Appropriate error estimators take care of the stiff and non-stiff components of the problems as to deliver a variable step size aiming at an integration error of the size of a tolerance given by the user. While efficient stabilized integrators for special regimes of (1.1) are available, none has existed until now for the various potential regimes of (1.1). We also emphasize that PIROCK is more than a simple combination of integrators developed in [VSH04, AL08, AVZ12, AVZ13b, Zbi11], as the coupling of the different regimes requires new ideas to stabilize the various possible combinations of the dynamics in (1.1).

[3] Notice also that optimal stabilized polynomial functions along the imaginary axis have only a linear growth with respect to the stage number [IV.2,1]. Thus, stabilized explicit integrators have no advantage for pure hyperbolic problems.

The PIROCK algorithm couples the standard ROCK2 integrator with the following classical deterministic methods. The noise terms are treated using similar ideas as for the S-SDIRK(2,2) and the S-ROCK2 methods.

$$
\begin{array}{c|ccc}
0 & & & \\
\frac{1}{3} & \frac{1}{3} & & \\
\frac{2}{3} & & \frac{2}{3} & \\
\hline
 & \frac{1}{4} & 0 & \frac{3}{4}
\end{array}
\qquad\qquad
\begin{array}{c|ccc}
\gamma & \gamma & & \\
1-\gamma & 1-2\gamma & \gamma & \\
\hline
 & \frac{1}{2} & \frac{1}{2} &
\end{array}
\tag{2.34}
$$

F_A–method (left) F_R–method (right)

where $\gamma = 1 - \sqrt{2}/2$. A 3-stage third order explicit method is taken for the advection (so that a non-empty portion $(-i\sqrt{3}, i\sqrt{3})$ of the imaginary axis is included in the stability domain of the F_A method) and a 2-stage second order singly diagonally implicit RK method for the reaction. This latter method is L-stable and can be efficiently implemented: due

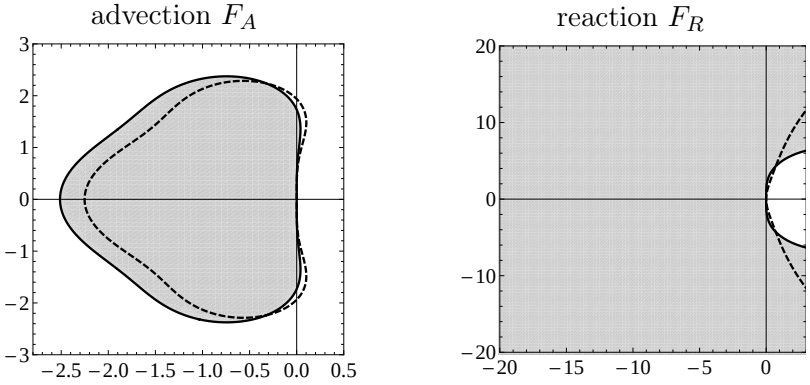


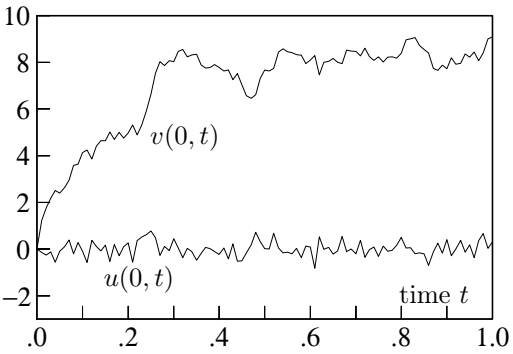
Figure 2.8: Complex stability domains (2.22) for the F_A and the F_R when applied to the linear test problem $dX(t)/dt = \lambda X(t)$. The dashed lines indicate the boundaries of the error estimator stability domains.

to the diagonal structure of the RK coefficients, a only single LU factorization needs to be done once per step if using a quasi-Newton method [HW96].

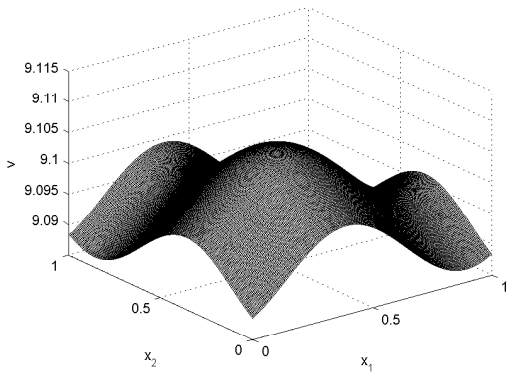
Example: a 2D brusselator with non-symmetric diffusion, advection, a highly stiff reaction, and stiff Itô stochastic noise To illustrate the versatility and efficiency of the proposed PIROCK integrator, we consider a modification of the Brusselator problem [HW96] with simultaneously all the difficulties of a non-symmetric diffusion operator, a stiff reaction, advection, and a two-dimensional stiff Itô stochastic noise, defined as

$$
\begin{aligned}
\frac{\partial u}{\partial t} &= \nu\Delta u + \nu/2\,\Delta v + \mu U\cdot\nabla u + \big(A + u^2 v - (B+1)u\big) + (\sigma_{11} + \sigma_{12}u)\dot{W}_1, \\
\frac{\partial v}{\partial t} &= -\nu/2\,\Delta u + \nu\Delta v + \big(\mu V\cdot\nabla v + f\big) + \big(Bu - u^2 v\big) + (\sigma_{21} + \sigma_{22}uv)\dot{W}_2. \qquad (2.35)
\end{aligned}
$$

For this problem we thus have to open all the blades of the "swiss-knife". For the diffusion and advection parameters, we take $\nu = 0.1, \mu = 0.1$, $U = (-0.5, 1)^T, V = (0.4, 0.7)^T$. We also consider a stiff reaction with parameters $A = 1.3, B = 10^7$, and with stiff noise parameters $\sigma_{11} = 3, \sigma_{12} = 4.4\cdot 10^3$, $\sigma_{21} = 0.5$, $\sigma_{22} = 1$. Since $-B + \sigma_{21}^2/2 < 0$, the reaction-noise system can be shown to be mean-square stable. We also consider an inhomogenity defined as $f(x) = 5$ if $(x_1 - 0.3)^2 + (x_2 - 0.6)^2 \le 0.3^2$, and $f(x) = 0$ else. We consider a space



(a) Solution at the spatial point $x = (0, 0)$.



(b) Solution $v(x, 1)$ at final time $t = 1$.

Figure 2.9: Non-symmetric diffusion-advection-reaction-noise problem (2.35). Space discretization: two 200×200 meshes. Constant step size $h = 10^{-2}$.

discretization with two 200×200 meshes and consider the constant time step size $h = 10^{-2}$ on the time interval $(0, 1)$. The number of stages used at each step to treat the diffusion is $s_{max} = 28$. We plot in Figure 2.9 one realisation of the problem (2.35). In picture 2.9(a), we plot the solutions $u(x, t), v(x, t)$ as a function of time t for $x = (0, 0)$ fixed, while in picture 2.9(b), we plot the solution $v(x, t)$ at final time $t = 1$ as a function of the spatial variable $x = (x_1, x_2)$. It can be seen that the solution oscillates stochastically in time, while it remains smooth in space. Note that for the standard Euler-Maruyama method, the step size restriction for mean-square stability can be estimated as $h \leq 0.64 \cdot 10^{-8}$, which makes this method of no practical use for this problem.

2.6 Perspectives

Note that the stabilization methodology used for deriving our weak order two integrators for stiff SDEs can in principle be adapted to stabilize higher order integrators. In addition, for a practical computation, the expectancy $\mathbb{E}(\phi(X_N))$ is approximated by a Monte-Carlo method [KP92]. The efficiency of this later approximation is not addressed in this work but very important in practice. In particular, the new weak order two integrators for stiff SDEs introduced could be combined with the recently proposed Multilevel Monte-Carlo method [Gil08].

It would be interesting the extend the swiss-knife integrator to treat the issues of constrained systems (e.g. the conservation of mass in the Navier-Stokes equation) or local spatial mesh refinements. This remains a challenge in the context of explicit stabilized integrators.

Chapter 3

Numerical homogenization methods for linear and nonlinear PDEs

This chapter is devoted to the numerical approximation of the solution of multiscale in space problems and summarizes the works [AV12a] and [AV11, AV12b, AV13a] in collaboration with A. Abdulle. A standard finite element method usually requires a very fine spatial mesh that resolves this multiscale structure. In contrast, an appropriate homogenization method can provide an accurate solution, but with a significantly reduced computational cost.

Multiscale media properties enter in the modeling of many important problems, we mention the infiltration of water in porous medium (e.g. the Richards problem [BB91]), and nonlinearities of the type considered in this chapter arise naturally in several applications, for instance the modeling of the thermal conductivity of the Earth's crust [WHN09], the study of electrical potential or thermal diffusion in composite materials [KL08] (see also the surveys [AEEVE12, Abd13] and reference therein).

3.1 Homogenization framework

Given a bounded, convex, and polyhedral domain Ω in $\mathbb{R}^d$, we focus successively on the following two classes of problems. Although they are of different natures, we would like to highlight in the chapter common phenomena and tools arising in their study:

- a class of linear parabolic problems with a multiscale time-dependent tensor,

$$\begin{aligned} &\partial_t u_\varepsilon - \nabla \cdot (a^\varepsilon(x,t)\nabla u_\varepsilon) = f \ \text{ in } \Omega \times (0,T) \\ &u_\varepsilon = 0 \ \text{ on } (0,T)\times \partial\Omega \\ &u_\varepsilon(x,0) = g(x) \ \text{ in } \ \Omega, \end{aligned} \tag{3.1}$$

  where $T>0$ is fixed and we assume $f \in L^2(0,T;L^2(\Omega))$, $g \in L^2(\Omega)$,

- and a class of nonlinear elliptic problems with a nonlinear multiscale tensor,

$$\begin{aligned} &-\nabla \cdot (a^\varepsilon(x,u_\varepsilon)\nabla u_\varepsilon) = f \ \text{ in } \Omega \\ &u_\varepsilon = 0 \ \text{ on } \partial\Omega, \end{aligned} \tag{3.2}$$

  where we assume $f \in H^{-1}(\Omega)$.

In (3.1), the tensor a^ε is time-dependent, while in (3.2) it depends non-linearly on the solution u_ε itself. In both cases, we assume that the tensor $a^\varepsilon(x,t)$ satisfies $a^\varepsilon \in (L^\infty(\Omega \times J))^{d\times d}$ where $J=(0,T)$ or $\mathbb{R}$, respectively, and that it is elliptic and bounded uniformly with respect to ε, i.e.,

$$\begin{aligned} &\exists \lambda, \Lambda > 0 \text{ such that } \lambda|\xi|^2 \le a^\varepsilon(x,t)\xi\cdot\xi, \ \ \|a^\varepsilon(x,t)\xi\| \le \Lambda\|\xi\|, \\ &\forall \xi \in \mathbb{R}^d \text{ and a.e. } x\in\Omega, t\in J, \forall \varepsilon>0. \end{aligned} \tag{3.3}$$

For the nonlinear problem (3.2), we make the additional assumption that $a^\varepsilon(x,s)$ is Lipschitz continuous with respect to s uniformly with respect to ε and a.e. x. Using the smoothness assumptions on the data, for all fixed $\varepsilon>0$, both problems (3.1) and (3.2) are known to have a unique solution u_ε in the Sobolev spaces[1]

$$E = \{v;\ v\in L^2(0,T;H_0^1(\Omega)),\ \partial_t v \in L^2(0,T;H^{-1}(\Omega))\} \tag{3.4}$$

for problem (3.1) (see for example [LM68]) and $E=H_0^1(\Omega)$ for problem (3.2) (see [Chi09, Thm. 11.6] for a proof), respectively. Using the assumptions (3.3) where all constants are

[1] The Sobolev space (3.4) is equipped with the usual norm $\|v\|_E = \|v\|_{L^2(0,T;H^1(\Omega))} + \|\partial_t v\|_{L^2(0,T;H^{-1}(\Omega))}$.

independent of ε, we have in both problems that u_ε is bounded in E uniformly with respect to ε. Standard compactness arguments then ensure the existence of a subsequence of $\{u_\varepsilon\}$ (still denoted by ε) such that

$$u_\varepsilon \rightharpoonup u_0 \text{ weakly in } E. \tag{3.5}$$

The aim of homogenization theory is to find and study a limiting equation for the weak limit u_0. Asymptotic expansions can be used to find a candidate for such a limiting equation. To show that the solution of this latter equation is the limit (in some sense) of the oscillating family of functions $\{u_\varepsilon\}$, one uses usually the notion of G or H convergence (see [Spa68, DGS73, MT97]), the former being restricted to symmetric tensors. It is then possible to show that there exists a subsequence of $\{u_\varepsilon\}$ (still denoted by ε) satisfying (3.5) and such that u_0 is the solution of a so-called homogenized problem of the same form as the multiscale problem (3.1)

$$\begin{aligned} &\partial_t u_0 - \nabla \cdot \left(a^0(x,t)\nabla u_0\right) = f \ \text{ in } \Omega \times (0,T) \\ &u_0 = 0 \ \text{ on } (0,T) \times \partial\Omega \\ &u_0(x,0) = g(x) \ \text{ in } \ \Omega, \end{aligned} \tag{3.6}$$

or the multiscale problem (3.2),

$$\begin{aligned} &-\nabla \cdot \left(a^0(x,u_0)\nabla u_0\right) = f \ \text{ in } \Omega \\ &u_0 = 0 \ \text{ on } \partial\Omega, \end{aligned} \tag{3.7}$$

respectively, with the exception that the tensor $a^\varepsilon(x,t)$ is replaced by an homogenized tensor $a^0(x,t)$ satisfying again (3.3), possibly with different constants λ, Λ. We refer to [BLP78, BOFM92, CD99] and [BM81] for details on homogenization theory in the context of linear parabolic problems and nonlinear elliptic problems, respectively.

Let us note that if $a^\varepsilon(x,t)$ has more structure as for example if $a^\varepsilon(x,t) = a(x,x/\varepsilon,t) = a(x,y,t)$ is periodic with respect to y, with $a(x,y+\mathbf{e}_i,t) = a(x,y,t)$ where $\mathbf{e}_i$, $i = 1,\dots,d$ denotes the canonical basis of $\mathbb{R}^d$, then one can show provided appropriate smoothness conditions that the whole sequence $\{u_\varepsilon\}$ weakly converges in the sense (3.5), without the need of extracting a subsequence. In addition, the homogenized tensor can be characterized in the following way [AD82]:

$$a^0(x,s) = \int_Y a(x,y,s)(I + J^T_{\chi(x,y,s)})dy, \qquad \text{for } x \in \Omega, s \in \mathbb{R}, \tag{3.8}$$

where $Y = (0,1)^d$, $J_{\chi(x,y,s)}$ is a $d \times d$ matrix with entries $J_{\chi(x,y,s)_{ij}} = (\partial\chi^i)/(\partial y_j)$ and $\chi^i(x,\cdot,s)$, $i = 1,\dots,d$ are the unique solutions of the cell problems

$$\int_Y a(x,y,s)\nabla_y\chi^i(x,y,s)\cdot\nabla w(y)dy = -\int_Y a(x,y,s)\mathbf{e}_i\cdot\nabla w(y)dy, \quad \forall w \in W^1_{per}(Y). \tag{3.9}$$

Remark 3.1.1 *We sometimes refer to the problems* (3.2) *or* (3.7) *as "non monotone problems". This stems from the following fact: writing for example* (3.7) *in weak form*

$$B(u_0;u_0,v) = \int_\Omega a^0(x,u_0(x))\nabla u_0(x)\nabla v(x)dx = (f,v), \quad \forall v \in H^1_0(\Omega)$$

we observe that the monotonicity property $B(u_0;u_0,u_0-v) - B(v;v,u_0-v) \geq C\|u_0 - v\|^2_{H^1\Omega}$ *with* $C \geq 0$ *does not hold in general for the quasilinear problem* (3.7). *This lack of monotonicity makes the numerical analysis for finite element method, specially when quadrature formula are used, a nontrivial task as shown in Section 3.4.1.*

Prohibitive cost of standard integrators The discretization of the problems (3.1) and (3.2) with a standard finite element method (FEM) is a well understood problem. Taking a finite dimensional subspace $S(\Omega, \mathcal{T}_h)$ of the Banach space (3.4), we search for a piecewise polynomial solution $u^h(t) \in S(\Omega, \mathcal{T}_h)$ of the variational equation corresponding to (3.1) in $S(\Omega, \mathcal{T}_h)$. However, the major issue is that usual convergence rates can only be obtained if the spatial meshsize h satisfies $h < \varepsilon$. For multiscale problems with order of magnitude of discrepancies between the scale of interest (for which we would like to set the spatial grid) and ε, the restriction $h < \varepsilon$ can be prohibitive in terms of degrees of freedom of the computational procedure if not impossible to realize.

A micro-macro homogenization method In this chapter, we focus on the so-called finite element Heterogeneous Multiscale Method (FE-HMM). The idea of the FE-HMM detailed in the next section is to rely on two grids, and in turn on two FE methods for the approximation of the homogenized solution u_0. A macroscopic method relying on a macroscopic mesh $H > \varepsilon$ which does not discretize the fine scale and microscopic meshes (defined on sampling domains within the macroscopic mesh) which discretize the smallest scale. Proper averaging of the microscopic FEM on the sampling domains allows to recover macroscopic (averaged) data related to the homogenized problem whose coefficients are unknown beforehand.

We mention that a popular alternative approach for multiscale PDE problems is the Multiscale Finite Element method (MsFEM) [HWC99, AB05] (see also [CS08] in the context of quasilinear problems of the type (3.2)), where the main idea is to enrich a coarse FE space with oscillating functions for the computation of the oscillatory solution u_ε. The computational complexity of the MsFEM is $\mathcal{O}(H^{-d}\varepsilon^{-d})$ where H denotes the mesh size and d is the dimension of the computational domain. Note that the MsFEM complexity grows as ε tends to zero. In contrast, denoting H and $\hat{h} = h/\varepsilon$ the macro mesh size and (scaled) micro mesh sizes, the complexity of the FE-HMM is $\mathcal{O}(H^{-d}\hat{h}^{-d})$ (consisting of $\mathcal{O}(H^{-d})$ independent linear micro problems with $\mathcal{O}(\hat{h}^{-d})$ degrees of freedom) and this is independent of the smallness of the parameter ε (optimal macro and micro mesh size refinement strategies are presented in Sect. 3.4.2.2).

Corrector procedure We emphasize that the FE-HMM aims at capturing the homogenized solution u^0 of (3.1) and (3.2). However, a numerical corrector can be defined extending the alreading computed micro solutions (defined in the sampling domains) on each whole macro element. With the help of a numerical corrector, an approximation of the fine scale solution u^ε can then be obtained (see [EMZ05],[ABDe09, Chap. 3.3.2] in the context of the linear FE-HMM). For nonlinear monotone elliptic problems, the convergence of such reconstruction procedure has been proved in [EP03, EP04] for the MsFEM in the stochastic case, and in [Glo06] for both the MsFEM and HMM in the general case of an arbitrary G-converging sequence.

3.2 The finite element heterogeneous multiscale method (FE-HMM)

In this section we describe the numerical method under study for the class of multiscale parabolic problems (3.1). It is based on the finite element heterogeneous multiscale methods, introduced an analyzed in [EE03, Abd05, EMZ05] (see [ABDe09, Abd11] for a review). The modifications of the method for solving the considered class of nonlinear problems (3.2)

are discussed in Section 3.4.2. In the HMM context, two approaches for the (spatial) numerical homogenization of parabolic problems have been proposed. The method in [AE03] is based on finite difference discretization techniques while the method in [MZ07] is based on finite element discretization techniques.

Macro finite element space We consider a partition $\mathcal{T}_H$ of Ω in simplicial or quadrilateral elements K of diameter H_K and denote $H := \max_{K\in\mathcal{T}_H} H_K$. We assume that this triangulation is conformal, shape regular. We consider the family of FE spaces

$$S_0^\ell(\Omega,\mathcal{T}_H) = \{v^H \in H_0^1(\Omega);\ v^H|_K \in \mathcal{R}^\ell(K),\ \forall K \in \mathcal{T}_H\}, \tag{3.10}$$

where $\mathcal{R}^\ell(K)$ is the space $\mathcal{P}^\ell(K)$ of polynomials on K of total degree at most ℓ if K is a simplicial FE, or the space $\mathcal{Q}^\ell(K)$ of polynomials on K of degree at most ℓ in each variable if K is a quadrilateral FE. We define a quadrature formula $\{\hat{x}_j,\hat{\omega}_j\}_{j=1}^J$ on a reference element $\hat{K}$, where $\hat{x}_j$ are integration points and $\hat{\omega}_j$ are quadrature weights. The quadrature formula $\{x_{j,K},\omega_{j,K}\}_{j=1}^J$ is then defined as usual on any element K of the triangulation using an affine transformation. We make the following assumptions, which are similar to the case of linear elliptic problems (see [CR72] or [Cia91, Sect. 29]):

(**Q1**) $\hat{\omega}_j > 0,\ j = 1,\dots,J$, and $\quad \sum_{j\in J}\hat{\omega}_j|\nabla\hat{p}(\hat{x}_j)|^2 \geq \hat{\lambda}\|\nabla\hat{p}\|^2_{L^2(\hat{K})},\ \forall\hat{p}(\hat{x}) \in \mathcal{R}^\ell(\hat{K})$;

(**Q2**) $\int_{\hat{K}}\hat{p}(x)dx = \sum_{j\in J}\hat{\omega}_j\hat{p}(\hat{x}_j),\ \forall\hat{p}(\hat{x}) \in \mathcal{R}^\sigma(\hat{K})$, where $\sigma = \max(2\ell-2,\ell)$ if $\hat{K}$ is a simplicial FE, or $\sigma = \max(2\ell-1,\ell+1)$ if $\hat{K}$ is a rectangular FE.

These requirements on the quadrature formula ensure that the optimal H^1 and L^2 convergence rates for standard FEM hold when using numerical integration [CR72].

Micro finite element spaces Based on the quadrature points, we define the microscopic sampling FE domains

$$K_{\delta_j} = x_{K_j} + \delta I,\ I = (-1/2,1/2)^d\ \ (\delta \geq \varepsilon). \tag{3.11}$$

We consider a (micro) partition $\mathcal{T}_h$ of each sampling domain K_{δ_j}, conformal and shape regular, in simplicial or quadrilateral elements Q of diameter h_Q and denote $h = \max_{Q\in\mathcal{T}_h} h_Q$. The sampling domains K_{δ_j} are typically of size ε, that is $|K_{\delta_j}| = \mathcal{O}(\varepsilon^d)$, and hence $h < \varepsilon \leq \delta$ holds for the micro mesh size. For this partition we define a micro FE space

$$S^q(K_{\delta_j},\mathcal{T}_h) = \{z^h \in W(K_{\delta_j});\ z^h|_T, \in \mathcal{R}^q(Q),\ Q \in \mathcal{T}_h\}, \tag{3.12}$$

where $W(K_{\delta_j})$ is a certain Sobolev space. Various spaces can be chosen for the micro numerical method, and the choice of the particular space has important consequences in the numerical accuracy of the method as we will see below (this choice sets the coupling condition between macro and micro solvers). We consider here

$$W(K_{\delta_j}) = W^1_{per}(K_{\delta_j}) = \{z \in H^1_{per}(K_{\delta_j});\ \int_{K_{\delta_j}} z dx = 0\}, \tag{3.13}$$

for a periodic coupling or

$$W(K_{\delta_j}) = H^1_0(K_{\delta_j}) \tag{3.14}$$

for a coupling through Dirichlet boundary conditions. Essential for the definition of the multiscale method below is the definition of the following microfunctions. Let $w^H \in S_0^\ell(\Omega,\mathcal{T}_H)$ and consider its linearization

$$w^H_{lin} = w^H(x_{K_j}) + (x - x_{K_j})\cdot\nabla w^H(x_{K_j})$$

at the integration point x_{K_j}. Associated to w_{lin}^H in the sampling domain K_{δ_j} we define a microfunction $w_{K_j}^{h,t}$, depending on t, such that $(w_{K_j}^{h,t} - w_{lin}^H) \in S^q(K_{\delta_j}, \mathcal{T}_h)$ and

$$\int_{K_{\delta_j}} a^\varepsilon(x,t)\nabla w_{K_j}^{h,t} \cdot \nabla z^h dx = 0 \quad \forall z^h \in S^q(K_{\delta_j}, \mathcal{T}_h). \tag{3.15}$$

We may now define the FE-HMM for (3.1).

Multiscale FE-HMM method. Find $u^H \in [0,T] \times S_0^\ell(\Omega, \mathcal{T}_H) \to \mathbb{R}$ such that

$$\begin{aligned} &(\partial_t u^H, v^H) + B_H(t; u^H, v^H) = (f(t), v^H) \ \ \forall v^H \in S_0^\ell(\Omega, \mathcal{T}_H) \\ &u^H = 0 \ \text{ on } \partial\Omega \times (0,T) \\ &u^H(x,0) = u_0^H, \end{aligned} \tag{3.16}$$

where

$$B_H(t; u^H, v^H) := \sum_{K \in \mathcal{T}_H} \sum_{j=1}^J \frac{\omega_{K_j}}{|K_{\delta_j}|} \int_{K_{\delta_j}} a^\varepsilon(x,t)\nabla u_{K_j}^{h,t} \cdot \nabla v_{K_j}^{h,t} dx, \tag{3.17}$$

and $u_0^H \in S_0^\ell(\Omega, \mathcal{T}_H)$ is chosen to approximate the exact initial condition g (see Remark 3.3.8 below). Here, $u_{K_j}^{h,t}, v_{K_j}^{h,t}$ are the solution of the microproblems (3.15) constrained by u_{lin}^H, v_{lin}^H, respectively.

3.3 Optimal a priori estimates for linear parabolic problems

3.3.1 Preliminaries: reformulation of the FE-HMM

In order to perform the analysis of the FE-HMM, it is convenient to introduce the following auxiliary bilinear form

$$B(t; v, w) = \int_\Omega a^0(x,t)\nabla v(x) \cdot \nabla w(x) dx, \quad \forall v, w \in H_0^1(\Omega), \tag{3.18}$$

where $a^0(x,t)$ is the homogenized tensor of (3.6). Consider also the associated bilinear form for standard FEM with numerical quadrature,

$$B_{0,H}(t; v^H, w^H) = \sum_{K \in \mathcal{T}_H} \sum_{j=1}^J \omega_{K_j} a^0(x_{K_j}, t)\nabla v^H(x_{K_j}) \cdot \nabla w^H(x_{K_j}), \tag{3.19}$$

for all $v^H, w^H \in S_0^\ell(\Omega, \mathcal{T}_H)$. Of course, $a^0(x,t)$ is usually unknown, otherwise there is no need for a multiscale method. In order to define the FEM with numerical quadrature (as in the above bilinear form) and for the analysis, some regularity on the tensor $a^0(x,t)$ is needed. We assume

(**H1**) $\quad a_{ij}^0, \partial_t a_{ij}^0 \in \mathcal{C}^0([0,T] \times \overline{K})$ for all $K \in \mathcal{T}_H$ for all $i,j = 1, \ldots, d$.

The following construction of a numerically homogenized tensor is useful (see [Abd11, Abd12] for details). Let $\mathbf{e}_i, \ i = 1, \ldots, d$, denote the canonical basis of $\mathbb{R}^d$. For each $\mathbf{e}_i$ and each $t \in [0,T]$, we consider the following elliptic problem

$$\int_{K_{\delta_j}} a^\varepsilon(x,t)\nabla \psi_{K_j}^{i,h,t}(x) \cdot \nabla z^h(x) dx = -\int_{K_{\delta_j}} a^\varepsilon(x,t)\mathbf{e}_i \cdot \nabla z^h(x) dx, \ \forall z^h \in S^q(K_{\delta_j}, \mathcal{T}_h), \tag{3.20}$$

where $S^q(K_{\delta_j}, \mathcal{T}_h)$ is defined in (3.12) with either periodic or Dirichlet boundary conditions. We also consider the problem

$$\int_{K_{\delta_j}} a^\varepsilon(x,t)\nabla\psi^{i,t}_{K_j}(x)\cdot\nabla z(x)dx = -\int_{K_{\delta_j}} a^\varepsilon(x,t)\mathbf{e}_i\cdot\nabla z(x)dx,\ \forall z\in W(K_{\delta_j}), \quad (3.21)$$

where the Sobolev space $W(K_{\delta_j})$ is defined in (3.13) or (3.14). We then define two tensors

$$a^0_K(x_{K_j},t) := \frac{1}{|K_{\delta_j}|}\int_{K_{\delta_j}} a^\varepsilon(x,t)\left(I + J^T_{\psi^{h,t}_{K_j}(x)}\right)dx, \quad (3.22)$$

where $J_{\psi^{h,t}_{K_j}(x)}$ is a $d\times d$ matrix with entries $\left(J_{\psi^{h,t}_{K_j}(x)}\right)_{i\ell} = (\partial\psi^{i,h,t}_{K_j})/(\partial x_\ell)$ and

$$\bar{a}^0_K(x_{K_j},t) := \frac{1}{|K_{\delta_j}|}\int_{K_{\delta_j}} a^\varepsilon(x,t)\left(I + J^T_{\psi^{t}_{K_j}(x)}\right)dx, \quad (3.23)$$

where $J_{\psi^{t}_{K_j}(x)}$ is a $d\times d$ matrix with entries $\left(J_{\psi^{t}_{K_j}(x)}\right)_{ik} = (\partial\psi^{i,t}_{K_j})/(\partial x_k)$.

Using the above numerically homogenized tensors (3.22) or (3.23) and the results of [Abd12] (see also Lemmas 11 and 12 of [Abd11]) we obtain the following reformulation of the bilinear form $B_H(\cdot,\cdot)$ of (3.17) which will be useful for the analysis.

Lemma 3.3.1 *The bilinear form $B_H(\cdot,\cdot)$ defined in* (3.17) *can be written as*

$$B_H(t;v^H,w^H) = \sum_{K\in\mathcal{T}_H}\sum_{j=1}^{J}\omega_{K_j}a^0_K(x_{K_j},t)\nabla v^H(x_{K_j})\cdot\nabla w^H(x_{K_j}). \quad (3.24)$$

Using (3.23) we can also define a bilinear form useful for the subsequent analysis

$$\bar{B}_H(t;v^H,w^H) = \sum_{K\in\mathcal{T}_H}\sum_{j=1}^{J}\omega_{K_j}\bar{a}^0_K(x_{K_j},t)\nabla v^H(x_{K_j})\cdot\nabla w^H(x_{K_j}). \quad (3.25)$$

Solving the parabolic problem (3.16) with the bilinear form $\bar{B}_H$ amounts to neglecting the micro errors, as the micro functions in (3.21) are exact.

We quantify the discrepancy between the bilinear forms $B_{0,H}(t;\cdot,\cdot)$ defined in (3.19) and $B_H(t;\cdot,\cdot)$ defined in (3.17). This will account for the error done at the microscale as well as the so-called modeling error, the error induced by artificial micro boundary conditions or non-optimal sampling of the micro structure. Consider the quantity

$$\begin{aligned} r_{HMM} &:= \sup_{K\in\mathcal{T}_H, x_{K_j}\in K, t\in[0,T]} \|a^0(x_{K_j},t) - a^0_K(x_{K_j},t)\|_F \\ &+ \sup_{K\in\mathcal{T}_H, x_{K_j}\in K, t\in[0,T]} \|\partial_t a^0(x_{K_j},t) - \partial_t a^0_K(x_{K_j},t)\|_F, \end{aligned} \quad (3.26)$$

where $\|\cdot\|_F$ denotes the Frobenius norm[2]. Following the strategy developed in [Abd12, Abd09, Abd11] for the error analysis, we can further decompose r_{HMM} into micro and

[2] The Frobenius norm of a matrix M is defined as $\|M\|_F = \sqrt{\text{trace}(M^TM)}$.

modeling error terms as

$$\begin{aligned} r_{HMM} &= \underbrace{\sum_{k=0}^{1} \sup_{K\in\mathcal{T}_H, x_{K_j}\in K, t\in[0,T]} \|\partial_t^k a^0(x_{K_j},t) - \partial_t^k \bar{a}_K^0(x_{K_j},t)\|_F}_{r_{MOD}} \\ &+ \underbrace{\sum_{k=0}^{1} \sup_{K\in\mathcal{T}_H, x_{K_j}\in K, t\in[0,T]} \|\partial_t^k \bar{a}_K^0(x_{K_j},t) - \partial_t^k a_K^0(x_{K_j},t)\|_F}_{r_{MIC}}, \end{aligned} \tag{3.27}$$

where we have used the tensor (3.23).

The following lemma is a consequence of the Cauchy-Schwarz inequality.

Lemma 3.3.2 *Let $B_{0,H}(t;\cdot,\cdot)$ and $B_H(t;\cdot,\cdot)$ be the bilinear forms defined in* (3.19) *and* (3.17)*, respectively. Then we have*

$$\begin{aligned} &|B_{0,H}(t;v^H,w^H) - B_H(t;v^H,w^H)| + |B'_{0,H}(t;v^H,w^H) - B'_H(t;v^H,w^H)| \\ &\le C r_{HMM} \|v^H\|_{H^1(\Omega)} \|w^H\|_{H^1(\Omega)} \end{aligned}$$

Analogously, the modeling and the micro error can be traced in the following lemma.

Lemma 3.3.3 *Let $B_{0,H}(t;\cdot,\cdot), B_H(t;\cdot,\cdot)$ and $\bar{B}_H(t;\cdot,\cdot)$ be the bilinear forms defined in* (3.19)*,* (3.17)*, and* (3.25)*, respectively. Then we have for all $v^H, w^H \in S_0^\ell(\Omega,\mathcal{T}_H)$,*

$$\begin{aligned} &|B_{0,H}(t;v^H,w^H) - \bar{B}_H(t;v^H,w^H)| + |B'_{0,H}(t;v^H,w^H) - \bar{B}'_H(t;v^H,w^H)| \\ &\le C r_{MOD} \|v^H\|_{H^1(\Omega)} \|w^H\|_{H^1(\Omega)}, \\ &|\bar{B}_H(t;v^H,w^H) - B_H(t;v^H,w^H)| + |\bar{B}'_H(t;v^H,w^H) - B'_H(t;v^H,w^H)| \\ &\le C r_{MIC} \|v^H\|_{H^1(\Omega)} \|w^H\|_{H^1(\Omega)}. \end{aligned}$$

Standard estimates for bilinear forms with numerical quadrature. Consider the usual nodal interpolant $\mathcal{I}_H : C^0(\overline{\Omega}) \to S_0^\ell(\Omega,\mathcal{T}_H)$ onto the FE space $S_0^\ell(\Omega,\mathcal{T}_H)$ defined in (3.10). The following estimates are based on the Bramble-Hilbert lemma and have first been derived in [CR72, Thm. 4 and Thm. 5]. They will often be used in our analysis. Assuming (**Q2**) and the regularity assumptions of Theorem 3.3.4 (see next section), we have for all $v^H, w^H \in S_0^\ell(\Omega,\mathcal{T}_H)$ (where $\mu = 0$ or 1),

$$|B(t;v^H,w^H) - B_{0,H}(t;v^H,w^H)| \le CH \|v^H\|_{H^1(\Omega)} \|w^H\|_{H^1(\Omega)}, \tag{3.28}$$

$$|B(t;\mathcal{I}_H u_0,w^H) - B_{0,H}(t;\mathcal{I}_H u_0,w^H)| \le CH^\ell \|u_0(t)\|_{W^{\ell+1,p}(\Omega)} \|w^H\|_{H^1(\Omega)}, \tag{3.29}$$

$$|B(t;\mathcal{I}_H u_0,w^H) - B_{0,H}(t;\mathcal{I}_H u_0,w^H)| \le CH^{\ell+\mu} \|u_0(t)\|_{W^{\ell+1,p}(\Omega)} \|w^H\|_{\bar{H}^2(\Omega)}. \tag{3.30}$$

where $\|w^H\|_{\bar{H}^2(\Omega)} = \big(\sum_{K\in\mathcal{T}_H} \|w^H\|^2_{H^2(K)}\big)^{1/2}$.

An α-accretive operator. For the time-discretization analysis, we introduce for each time t the linear operator $A_H(t) : S_0^\ell(\Omega,\mathcal{T}_H) \to S_0^\ell(\Omega,\mathcal{T}_H)$ defined as

$$(-A_H(t)v^H, w^H) = B_H(t;v^H,w^H), \quad \text{for all } v^H, w^H \in S_0^\ell(\Omega,\mathcal{T}_H), \tag{3.31}$$

where B_H is the bilinear form defined in (3.17). Consider the sector in the complex plane

$$S_\alpha = \{\rho e^{i\theta} \ ; \ \rho \ge 0, |\theta| \le \alpha\}.$$

The operator A_H can be extended straightforwardly to a complex Hilbert space based on $S_0^\ell(\Omega, \mathcal{T}_H)$ equipped with the complex scalar product $(u, v) = \int_\Omega u(x)\overline{v}(x)dx$ which is an extension of the usual L^2 scalar product. It can be shown that $-A_H$ is a so-called α-accretive operator[3]: there exist $0 \leq \alpha \leq \pi/2$ and $C > 0$ such that for all $z \notin S_\alpha$, the operator $zI + A_H(t)$ is an isomorphism and

$$\|(zI + A_H(t))^{-1}\|_{L^2(\Omega)\to L^2(\Omega)} \leq \frac{1}{d(z, S_\alpha)} \qquad \text{for all } z \notin S_\alpha, \tag{3.32}$$

where $d(z, S_\alpha)$ is the distance between z and S_α. In general one can show that $\alpha \in [0, \pi/2)$ using the ellipticity and boundedness of the tensor. In the case of a symmetric tensor, all the eigenvalues of A_H are real and located on the negative real axis of the complex plane, and one has simply $\alpha = 0$. The proof of (3.32) is omitted as this is a classical result for the time discretization of parabolic PDEs. More details can be found for instance in [Cro05].

3.3.2 Fully-discrete analysis of the multiscale spatial discretization for a time-dependent tensor

We main now state the main result of this section.

Theorem 3.3.4 *Consider u_0, u^H the solutions of* (3.6), (3.16)*, respectively. Let $\mu = 0$ or* 1*, $\ell \geq 1$ and $2 \leq p \leq \infty$ such that $\ell > d/p$. Assume (***Q1***),(***Q2***),(***H1***),* (3.3) *and*

$$\begin{aligned} &u_0, \partial_t u_0 \in L^2(0, T; W^{\ell+1,p}(\Omega)), \\ &a^0_{ij}, \partial_t a^0_{ij} \in L^\infty(0, T; W^{\ell+\mu,\infty}(\Omega)), \quad \forall i, j = 1 \dots d. \end{aligned}$$

Then we have the $L^2(H^1)$ and $\mathcal{C}^0(L^2)$ estimates

$$\|u_0 - u^H\|_{L^2([0,T];H^1(\Omega))} \leq C(H^\ell + r_{HMM} + \|g - u_0^H\|_{L^2(\Omega)}), \tag{3.33}$$

$$\|u_0 - u^H\|_{\mathcal{C}^0([0,T];L^2(\Omega))} \leq C(H^{\ell+1} + r_{HMM} + \|g - u_0^H\|_{L^2(\Omega)}), \quad \text{if } \mu = 1. \tag{3.34}$$

If in addition, the tensor is symmetric, then we have the $\mathcal{C}^0(H^1)$ estimate

$$\|u_0 - u^H\|_{\mathcal{C}^0([0,T];H^1(\Omega))} \leq C(H^\ell + r_{HMM} + \|g - u_0^H\|_{H^1(\Omega)}). \tag{3.35}$$

The constants C are independent of H, r_{HMM}.

The first term in the right-hand side of the above estimates quantifies the error of the macro solver. It shows that the proposed multiscale FEM gives optimal (macroscopic) convergence rates in the $\mathcal{C}^0(L^2)$ and $L^2(H^1)$ norms (and $\mathcal{C}^0(H^1)$ for symmetric tensors) of the fully discrete FE-HMM (3.16). We emphasize that the above error estimates have been derived without specific assumptions on the oscillation of the multiscale tensor. We recall that the additional term r_{HMM} defined in (3.26), that appears in the right-hand side of (3.33) or (3.34), encodes the so-called modeling and micro error, i.e., the error due to a possible mismatch of the averaging procedure in the FE-HMM, the boundary conditions and size of the sampling domains as well as the discretization error done of the micro FEMs.

To quantify further the term r_{HMM} we need some regularity and growth assumption (in terms of ε) of the solution of the microproblems (3.21). Motivated by the case of periodic tensors (e.g. the chain rule applied to $a^\varepsilon = a(x, x/\varepsilon, t)$) we consider the following regularity

[3] Equivalently, $+A_H$ is called an α-dissipative operator.

assumption on the solution of problem (3.21)

(**H2**) $|\psi^{i,t}_{K_j}|_{H^{q+1}(K_{\delta_j})} + |\partial_t \psi^{i,t}_{K_j}|_{H^{q+1}(K_{\delta_j})} \leq C\ \varepsilon^{-q}\sqrt{|K_{\delta_j}|}$, where C is independent of ε, the time t, the quadrature points x_{K_j}, and the domain K_{δ_j}. We also suppose that the map $t \to a^\varepsilon(\cdot,t) \in (L^\infty(\Omega))^{d\times d}$ is C^1 and $|\partial_t a^\varepsilon_{ij}(t,\cdot)|_{L^\infty(\Omega)} \leq C$, for all $t \in (0,T)$ and all $\varepsilon > 0$. We make the same assumptions on the solution of the modified problem (3.21) where the tensor a^ε is replaced by $a^{\varepsilon T}$ (the adjoint problem).

Remark 3.3.5 *When Dirichlet boundary conditions* (3.14) *are imposed in* (3.20)*, the assumption (***H2***) can be easily satisfied (without any further knowledge about the structure of the oscillating tensor a^ε) for $q = 1$ as $|\psi^{i,t}_{K_j}|_{H^2(K_{\delta_j})} \leq C\varepsilon^{-1}\sqrt{|K_{\delta_j}|}$ follows from classical H^2 regularity results ([Lad85, Chap. 2.6]), provided that $|a^\varepsilon_{ij}(\cdot,t)|_{W^{1,\infty}(\Omega)} \leq C\varepsilon^{-1}$ for $i,j = 1,\dots,d$. Then, following the proof of [AV13a, Lemma 4.12], $|\partial_t \psi^{i,t}_{K_j}|_{H^2(K_{\delta_j})} \leq C\varepsilon^{-1}\sqrt{|K_{\delta_j}|}$ holds, provided $|\partial_t a^\varepsilon_{ij}(\cdot,t)|_{W^{1,\infty}(\Omega)} \leq C\varepsilon^{-1}$. For periodic boundary conditions* (3.13) *in* (3.20)*, (***H2***) can be established for any given q, provided $a^\varepsilon = a(x,x/\varepsilon,t) = a(x,y,t)$ is Y-periodic in y, $\delta/\varepsilon \in \mathbb{N}$, and a^ε is sufficiently smooth, by following classical regularity results for periodic problems (see [BJS64, Chap. 3]).* [4]

Using the smoothness assumption (**H2**) permits to estimate the quantity r_{MIC} in (3.27), while the following structure assumption of a periodic tensor permits to estimate r_{MOD}.

(**H3**) $a^\varepsilon = a(x,x/\varepsilon,t) = a(x,y,t)$ $\quad Y$-periodic in y, where we set $Y = (0,1)^d$.

We then have the following theorem.

Theorem 3.3.6 *Consider u_0, u^H the solutions of* (3.6), (3.16)*, respectively. In addition to the assumptions of Theorem 3.3.4, assume (***H2***) and (***H3***). Assume also that $\psi^{i,t}_{K_{\delta_j}}$ is the solution of the cell problem (3.20) in the space $W^1_{per}(K_{\delta_j})$, that $\varepsilon/\delta \in \mathbb{N}$, and that the tensor $a(x,x/\varepsilon,t)$ is collocated at the quadrature points $a(x_{K_j},x/\varepsilon,t)$ in the FE-HMM macro bilinear form* (3.17) *and in the micro problems* (3.15)*. Then we have*

$$\begin{aligned}
\|u_0 - u^H\|_{L^2([0,T];H^1(\Omega))} &\leq C(H^\ell + \left(\frac{h}{\varepsilon}\right)^{2q} + \|g - u_0^H\|_{L^2(\Omega)}),\\
\|u_0 - u^H\|_{\mathcal{C}^0([0,T];L^2(\Omega))} &\leq C(H^{\ell+1} + \left(\frac{h}{\varepsilon}\right)^{2q} + \|g - u_0^H\|_{L^2(\Omega)}), \quad \text{if } \mu = 1. \quad (3.36)
\end{aligned}$$

If in addition, the tensor is symmetric, then

$$\|u_0 - u^H\|_{\mathcal{C}^0([0,T];H^1(\Omega))} \leq C(H^\ell + \left(\frac{h}{\varepsilon}\right)^{2q} + \|g - u_0^H\|_{H^1(\Omega)}).$$

The constants C are independent of H, h, ε.

The first term in Theorem 3.3.6 quantifies the error coming from the macro solver. The second term quantities the error coming from the micro solver – when discretizing the microproblems by a FEM – transmitted to macroscale. This term does not appear in the analysis given in [MZ07], where the microsolutions u^h, v^h in (3.17) where supposed to be exact. The additional analysis of the micro error allows to derive a macro-micro refinement strategy.

[4] We also note that $\partial_t^k a^\varepsilon_{ij}|_K \in W^{1,\infty}(K)$ $\forall K \in \mathcal{T}_H$ and $|\partial_t^k a^\varepsilon_{ij}|_{W^{1,\infty}(K)} \leq C\varepsilon^{-1}$ with $k = 0$ and 1 are sufficient, if the macro mesh is aligned with the (possible) discontinuities of a^ε (see [Abd12] for details).

Remark 3.3.7 *We emphasize that the remaining term r_{MOD} defined in* (3.27) *does not depend on the macro and micro mesh sizes H and h. In particular, any result concerning the approximation of the homogenized tensor with artificial micro boundary conditions or modified cell problems (e.g. [BP04],[EMZ05],[BB10],[Glo11],[Glo12],[Yur86]) could be used in our analysis. If the tensor $a(x, x/\varepsilon, t)$ is not collocated at the slow variable in the above theorem, we get for the modeling error (see [AV13a, Appendix],[ABDe09, Prop. 14])*

$$r_{MOD} \leq C\ \delta.$$

If the solution of the cell problem (3.20) in $H_0^1(K_{\delta_j})$, a resonance error contributes to r_{MOD}. For a tensor independent of time, the results in [EMZ05] can be readily used in the framework developed here for the analysis of parabolic problems and we have

$$r_{MOD} \leq C(\delta + \frac{\varepsilon}{\delta}).$$

This results could be extended for time-dependent tensor by following [EMZ05] and [AV13a, Appendix].

3.3.3 Coupling with strongly A-stable implicit Runge-Kutta methods

In this section, we explain how fully discrete estimates in both space and time can be derived. We focus on on implicit time discretizations (Runge-Kutta methods) with variable timesteps analyzed in [LO95]. The case of explicit stabilized integrators (Chebyshev methods) is investigated in the next Section. We assume that the numerical initial condition u_0^H of the FE-HMM in (3.16) is chosen to approximate the exact initial condition g as

$$\|u_0^H - g\|_{L^2(\Omega)} \leq C(H^{\ell+1} + r_{HMM}), \tag{3.37}$$

$$\|u_0^H - g\|_{H^1(\Omega)} \leq C(H^{\ell} + r_{HMM}). \tag{3.38}$$

Remark 3.3.8 *There are several natural choices for the initial condition u_0^H to satisfy* (3.37)-(3.38). *For instance, one can take $u_0^H = \Pi_H g$, the L^2 projection of g on $S_0^\ell(\Omega, \mathcal{T}_H)$, defined as*

$$(\Pi_H g - g, z^H) = 0, \quad \forall z^H \in S_0^\ell(\Omega, \mathcal{T}_H), \tag{3.39}$$

and then (3.37)-(3.38) *hold without the r_{HMM} terms*[5]. *One can also consider the elliptic projection $u_0^H = P_H g$ with respect to the bilinear forms B in* (3.18) *and B_H in* (3.17),

$$B_H(0; P_H g, z^H) = B(0; g, z^H), \qquad \forall z^H \in S_0^\ell(\Omega, \mathcal{T}_H), \tag{3.40}$$

and (3.37)-(3.38) *hold.*

We consider a subclass of Runge-Kutta methods with coefficients $a_{ij}, b_j, j = 1, \ldots, s$ which are of order r with stage order (the accuracy of the internal stages) $r-1$, and which are strongly $A(\theta)$-stable with $0 \leq \theta \leq \pi/2$. This latter condition means that $I - z\Gamma$ (where $\Gamma = (a_{ij})$) is a nonsingular matrix in the sector $|\arg(-z)| \leq \theta$ and the stability function [6] $R(z) = 1 + zb^T(I - z\Gamma)^{-1}1$ satisfies $|R(z)| < 1$ in $|\arg(-z)| \leq \theta$ (we refer to [HW96, Sect. IV.3,IV.15] for details on the stability concepts described here).

[5] Note that the regularity assumed on $u_0, \partial_t u_0$ in Theorem 3.3.4 implies $u_0(0) = g \in W^{\ell+1,p}(\Omega)$.

[6] We recall that the stability function of a Runge-Kutta method is the rational function $R(\Delta t\lambda) = R(z)$ obtained after applying the method over one step Δt to the scalar problem $dy/dt = \lambda y,\ y(0) = 1,\ \lambda \in \mathbb{C}$.

Note that all s-stage Radau Runge-Kutta methods satisfy the above assumptions (with $\theta = \pi/2$) [HW96]. In particular, for $s = 1$, we retrieve the implicit Euler method

$$(M + \Delta t K(t_{n+1}))U^H_{n+1} = MU^H_n + F_H(t_{n+1}). \tag{3.41}$$

where M denotes the mass matrix and $K(t)$ denotes the stiffness matrix associated to the FE-HMM bilinear form (3.17) in the basis of the macro FE space $S^\ell_0(\Omega, \mathcal{T}_H)$. Our analysis for implicit methods covers variable time step methods, provided that the stepsize sequence $\{\Delta t_n\}_{0\le n\le N-1}$ with $\Delta t_n = t_{n+1} - t_n > 0$ and $t_N = T$ satisfies for $C, c > 0$

$$\sum_{n=0}^{N-1} |\Delta t_{n+1}/\Delta t_n - 1| \le C, \qquad c\Delta t \le \Delta t_n \le \Delta t \text{ for all } 0 \le n \le N-1. \tag{3.42}$$

The condition (3.42) may appear restrictive. However, a finite subdivision of the interval $[0, T]$ into subintervals can be considered and (3.42) is required only on each of the subintervals (see [LO95, Sect. 5]). This permits to use stepsizes of different scales.

The first theorem treats the case of implicit methods and is obtained by combining our fully discrete error estimates in space (Theorem 3.3.6) with the results of [LO95].

Theorem 3.3.9 *Consider u_0 the exact solution of (3.6) and u^H_n the numerical solution of a Runge-Kutta method for the FE-HMM problem (3.16), with variable timesteps $\{\Delta t_n\}$ satisfying (3.42). Given an integer $r \ge 1$, assume that the Runge-Kutta method has order r when applied to ordinary differential equations, that it has stage order $r - 1$, and that it is strongly $A(\theta)$-stable with $\alpha < \theta$ where α is the angle in (3.32) of accretivity of $-A_H$. Assume the hypotheses of Theorem 3.3.6 with $\mu = 1$. Assume further (3.37),*

$$f \in H^r(0, T; L^2(\Omega)), \quad a^\varepsilon \in \mathcal{C}^r([0, T], L^\infty(\Omega)^{d\times d}) \text{ with } \|\partial^k_t a^\varepsilon\|_{(L^\infty(\Omega))^{d\times d}} \le C, \ k = 1 \dots r,$$

and

$$\|\partial^r_t u^H(0)\|_{L^2(\Omega)} \le C, \tag{3.43}$$

where u^H is the solution of (3.16). Then, we have the $\mathcal{C}^0(L^2)$ estimate

$$\max_{0\le n\le N} \|u^H_n - u_0(t_n)\|_{L^2(\Omega)} \le C\left(H^{\ell+1} + \left(\frac{h}{\varepsilon}\right)^{2q} + \Delta t^r\right). \tag{3.44}$$

Assuming in addition (3.38) and that a^ε is symmetric, then we have the $L^2(H^1)$ estimate

$$\sum_{n=0}^{N-1} \Delta t_n \|u^H_n - u_0(t_n)\|^2_{H^1(\Omega)} \le C\left(H^\ell + \left(\frac{h}{\varepsilon}\right)^{2q} + \Delta t^r\right)^2. \tag{3.45}$$

All the above constants C are independent of $H, h, \varepsilon, \Delta t$.

The assumption (3.43) can be satisfied in dimension $d = \dim\Omega \le 3$ as proved in [AV12a, Prop. 5.3] . We mention that for $r = 1$ the symmetry assumption on the tensor can be removed for the estimate (3.45).

3.3.4 Coupling with explicit stabilized time-integrators

Recall form Section 2.4 that Chebyshev methods are a subclass of explicit Runge-Kutta methods with extended stability domains along the negative real axis, which make them integrators of choice for diffusion problems as considered in this chapter.

In [VHS90], convergence rates in time independent of the spatial discretization parameters have been derived for a class of linear parabolic problems for the RKC method [SSV98]. In this section, we extend such analysis to classes of explicit stabilized methods (including ROCK2) in our context of multiscale homogenization problems.

We focus for simplicity on the case where the tensor a^ε is symmetric and time-independent. Recall that it is essential when considering Chebychev methods that the eigenvalues of the differential operator of the problem remain close to the negative real axis. This is automatically the case when the tensor is symmetric.

Chebyshev methods are usually used in a "damped" form, where the stability function satisfies the strong stability condition

$$\sup_{z\in[-L_s,-\gamma],\, s\geq 1} |R_s(z)| < 1, \quad \text{for all } \gamma > 0, \tag{3.46}$$

where L_s denotes the length of the stability domain, which grows quadratically with respect to the stage number s (related to the number of diffusion function evaluations). For the analysis, we shall also need that the stability functions are bounded in a neighbourhood of zero uniformly with respect to s, precisely, there exist $\gamma > 0$ and $C > 0$ such that[7]

$$|R_s(z)| \leq C \text{ for all } |z| \leq \gamma \text{ and all } s. \tag{3.47}$$

Theorem 3.3.10 *Consider u_0 the exact solution of* (3.6) *and u_n^H the numerical solution of a Chebyshev method for the FE-HMM problem* (3.16)*, applied with a constant timestep $\Delta t = T/N$, and with stability functions $\{R_s(z)\}_{s\geq 1}$. Assume that the tensor a^ε is symmetric and time-independent, and that $f = 0$. Assume* (3.37) *and the hypotheses of Theorem 3.3.6 with $\mu = 1$. Given $r \geq 1$, assume that the order of the Chebyshev method is r, precisely,*

$$\lim_{z\to 0} \left| \frac{e^z - R_s(z)}{z^{r+1}} \right| < \infty \quad \text{for all } s \geq 1. \tag{3.48}$$

In addition to (3.47)*, assume the strong stability condition* (3.46) *holds with the number of stages s chosen such that $\rho\Delta t \leq L_s$, where ρ is the spectral radius of the operator A_H defined in* (3.31)*. Then,*

$$\max_{0\leq n\leq N} \|u_n^H - u_0(t_n)\|_{L^2(\Omega)} \leq C \left(H^{\ell+1} + \left(\frac{h}{\varepsilon}\right)^{2q} + \Delta t^r \right). \tag{3.49}$$

For the sake of brevity of the analysis, we assumed in Theorem 3.3.10 above that the source term f is zero. Note that a non-zero time-independent source $f(x)$ could also be considered in the analysis by using a change of variable of the standard form $u_0(x,t) \leftrightarrow u_0(x,t) - \overline{u}_0(x)$ where $\overline{u}_0$ denotes the stationary solution of the problem, to retreive the zero source case (we omit the details). Moreover, in the case where the strong stability condition (3.46) is not satisfied (for instance if the damping is zero in the Chebyshev method (2.20)), we can still show the convergence by exploiting the regularity of the initial condition, as illustrated in [AV12a, Thm. 5.6].

Remark 3.3.11 *For simplicity, we assumed $r_{MOD} = 0$ in Theorems 3.3.9 and 3.3.10. If (***H3***) does not hold, then* (3.44)*,* (3.45)*, and* (3.49) *remain valid provided the term r_{MOD} defined in* (3.27) *is added in the right-hand sides of these estimates.*

[7] The estimate (3.47) can be easily checked for the the Chebyshev methods (2.20) using the standard formula $T_s(x) = (\xi_1^s + \xi_2^s)/2$ where ξ_1, ξ_2 are the complex roots of $X^2 - 2xX + 1$.

3.4 Optimal a priori estimates for nonlinear non-monotone elliptic problems

The finite element heterogeneous multiscale method relies of the standard finite element method (FEM) with numerical quadrature. We first derive in Section 3.4.1 convergence estimates for the standard FEM with the numerical quadrature, and then derive in Section 3.4.1.1 the convergence analysis of the FE-HMM.

3.4.1 The one scale case: analysis of numerical quadrature effects in standard nonlinear finite element methods

We study finite element (FE) discretizations of second-order quasilinear elliptic problems of the form (3.7),

$$-\nabla\cdot(a(x,u(x))\nabla u(x)) = f(x) \ \text{ in } \Omega, \qquad u(x)=0 \ \text{ on } \partial\Omega, \tag{3.50}$$

where Ω is a bounded convex polyhedron in $\mathbb{R}^d$ with $d\leq 3$. We recall the assumptions made on the tensor $a(x,s)=(a_{mn}(x,s))_{1\leq m,n\leq d}$:

- the coefficients $a_{mn}(x,s)$ are continuous functions on $\overline{\Omega}\times\mathbb{R}$ which are uniformly Lipschitz continuous with respect to s, i.e.,

$$\exists\Lambda_1>0, \quad |a_{mn}(x,s_1)-a_{mn}(x,s_2)|\leq\Lambda_1|s_1-s_2|,\ \forall x\in\overline{\Omega},\forall s_1,s_2\in\mathbb{R},\forall\, 1\leq m,n\leq d. \tag{3.51}$$

- $a(x,s)$ is uniformly coercive and bounded, i.e.,

$$\exists\lambda,\Lambda_0>0, \quad \lambda\|\xi\|^2\leq a(x,s)\xi\cdot\xi, \qquad \|a(x,s)\xi\|\leq\Lambda_0\|\xi\|, \qquad \forall\xi\in\mathbb{R}^d,\forall x\in\overline{\Omega},\forall s\in\mathbb{R}. \tag{3.52}$$

Since (3.51)-(3.52) hold, it is known (see e.g. [Chi09, Thm. 11.6]) that (3.50) has a unique solution $u\in H^1_0(\Omega)$ for all $f\in L^2(\Omega)$.

The convergence in $H^1(\Omega)$ of the FE solution with numerical quadrature was first shown in [FKS93] for piecewise linear FEs, without convergence rates. Note that the differential operator associated to (3.50) is not monotone in general, so the analysis in [FŽ87] does not apply here. In the absence of numerical quadrature, optimal a priori error estimates in the H^1 and L^2 norms for FE methods (FEMs) were first given in [DD75]. The case of a FEM with numerical quadrature is considered in this section. As exact integration in FEMs is rarely possible, it is important to quantify the effect of numerical quadrature. This is an essential ingredient of the analysis of the nonlinear FE-HMM conducted in Section 3.4.2. Optimal convergence rates in the H^1 and L^2 norms are proved in this case. The practical implementation of the non-linear FEM requires a Newton method. We also establish the convergence of this latter method (crucial in applications) and the uniqueness of the FE solution for a sufficiently fine FE mesh. If $a(x,s)$ becomes independent of s, we recover the results of [CR72] on FEMs with numerical quadrature for linear problems (convex polyhedral domain case).

3.4.1.1 Finite element method with numerical quadrature

Consider the partition $\mathcal{T}_H$ of Ω in simplicial or quadrilateral elements K satisfying the usual assumptions (see Section 3.2). For the nonlinear analysis, we make the additional assumption of a quasi-uniform mesh,

$$\frac{H}{H_K}\leq C \ \text{ for all } K\in\mathcal{T}_H \text{ and all } \mathcal{T}_H \text{ of the family of triangulations.} \tag{3.53}$$

Consider for v, w scalar or vector functions that are piecewise continuous with respect to the partition $\mathcal{T}_H$ of Ω, the semi-definite inner product

$$(u, v)_H := \sum_{K \in \mathcal{T}_H} \sum_{j=1}^{J} \omega_{j,K} u(x_{j,K}) v(x_{j,K}).$$

The FE solution of (3.50) with numerical integration reads: find $u^H \in S_0^\ell(\Omega, \mathcal{T}_H)$ such that

$$(a(\cdot, u^H)\nabla u^H, \nabla w^H)_H = F_H(w^H) \qquad \forall w^H \in S_0^\ell(\Omega, \mathcal{T}_H), \tag{3.54}$$

where the linear form $F_H(w^H)$ is an approximation of $\int_\Omega f(x) w^H(x) dx$ obtained for example by using a quadrature formula. If $f \in W^{\ell,q}(\Omega)$ with $1 \le q \le \infty$ and $\ell > d/q$, then f is continuous on $\overline{\Omega}$ and one can take $F_H(w^H) := (f, w^H)_H$.

The existence of the FE solution $u^H \in S_0^\ell(\Omega, \mathcal{T}_H)$ in (3.54) can be shown for all $H > 0$ using the Brouwer fixed point theorem. Details can be found for example in [DD75] (see also [BS08]).

3.4.1.2 A priori error analysis for non-monotone problems

The following theorem states the that the convergence rates in the H^1 and L^2 norms of standard FEM with numerical quadrature for the class of nonlinear elliptic problems (3.50) are identical to the linear elliptic case.

Theorem 3.4.1 *Consider u the solution of problem (3.50). Let $\ell \ge 1$. Let $\mu = 0$ or 1. Assume (**Q1**), (**Q2**), (3.53), and*

$$\begin{array}{ll} u \in H^{\ell+1}(\Omega) \cap W^{1,\infty}(\Omega), & \\ a_{mn} \in W^{\ell+\mu,\infty}(\Omega \times \mathbb{R}), & \forall m, n = 1 \dots d, \\ f \in W^{\ell+\mu,q}(\Omega), & \text{where } 1 \le q \le \infty, \ell > d/q. \end{array}$$

In addition to (3.51), (3.52), assume that $\partial_u a_{mn} \in W^{1,\infty}(\Omega \times \mathbb{R})$, and that the coefficients $a_{mn}(x, s)$ are twice differentiable with respect to s, with the first and second order derivatives continuous and bounded on $\overline{\Omega} \times \mathbb{R}$, for all $m, n = 1 \dots d$.[8]

Then there exists $H_0 > 0$ such that for all $H \le H_0$, the solution u^H of (3.54) is unique, and the following H^1 and L^2 error estimates hold,

$$\text{if } \mu = 0, 1, \qquad \|u - u^H\|_{H^1(\Omega)} \le C h^\ell \qquad \text{for all } h \le h_0, \tag{3.55}$$

$$\text{if } \mu = 1, \qquad \|u - u^H\|_{L^2(\Omega)} \le C h^{\ell+1} \qquad \text{for all } h \le h_0, \tag{3.56}$$

where the constant C is independent of h.

Inspired by [DD75], the proof of Theorem 3.4.1 is conducted in three main steps.
Step 1. Using the compact injection $H^1(\Omega) \subset L^2(\Omega)$, the boundedness of a numerical solution in $H_0^1(\Omega)$ and the uniqueness in $H_0^1(\Omega)$ of the exact solution of (3.50), we show,

$$\|u - u^H\|_{L^2(\Omega)} \to 0 \text{ for } H \to 0. \tag{3.57}$$

Step 2. We derive the following H^1 a priori error bound

$$\|u - u^H\|_{H^1(\Omega)} \le C(H^\ell + \|u - u^H\|_{L^2(\Omega)}), \text{ for all } H > 0. \tag{3.58}$$

[8] The uniqueness of u^H for all $H \le H_0$ and the H^1 estimate (3.55) for all $H > 0$ both hold without this additional assumption and without (3.53) and $u \in W^{1,\infty}(\Omega)$, if $C\lambda^{-1}\Lambda_1 \|u\|_{H^2(\Omega)} < 1$ (where C depends only on Ω and $(S_0^\ell(\Omega, \mathcal{T}_H))_{H>0}$).

Compared to the linear case, the additional term $\|u-u^H\|_{L^2(\Omega)}$ in the right-hand side is due to the non-monotonicity of the differential operator of (3.50). The proof of (3.58) relies on an estimate for $(a(u^H)\nabla u^H, \nabla w^H) - (a(u^H)\nabla u^H, \nabla w^H)_H$ (obtained by using the Bramble-Hilbert lemma and generalizing to a nonlinear context the estimates (3.28)(3.29)(3.30)), and the use of the Gagliardo-Niremberg inequality $\|v\|^2_{L^3(\Omega)} \leq C\|v\|_{L^2(\Omega)}\|v\|_{H^1(\Omega)}$, which holds for all $v \in H^1(\Omega)$ for $d \leq 3$.

Step 3. We show that there exists $H_1 > 0$ such that

$$\|u - u^H\|_{L^2(\Omega)} \leq C(H^{\ell+\mu} + \|u - u^H\|^2_{H^1(\Omega)}), \text{ for all } H \leq H_1. \tag{3.59}$$

This estimates relies on an Aubin-Nitsche duality argument, where we consider the adjoint L^* of the linearized operator associated to (3.50),

$$L\varphi := -\nabla \cdot (a(\cdot, u)\nabla\varphi + \varphi\partial_u a(\cdot, u)\nabla u). \tag{3.60}$$

An estimate between the FE solution with numerical quadrature of (3.60) and its exact solution is crucial. It is established by using estimates for FEM with numerical quadrature for indefinite linear elliptic problems (using a variant of Proposition 3.4.2 below).

Proof. [Proof of the H^1 and L^2 estimates.] Substituting (3.58) into (3.59) (with $\mu = 0$), we obtain

$$\|u - u^H\|_{H^1(\Omega)} \leq C(H^{\ell} + \|u - u^H\|^2_{H^1(\Omega)}), \text{ for all } H \leq H_1.$$

Substituting (3.57) into (3.58), we obtain $\|u - u^H\|_{H^1(\Omega)} \to 0$ for $h \to 0$. We deduce in the above inequality $1 - C\|u - u^H\|_{H^1(\Omega)} \geq \delta > 0$ for all $h \leq h_2$, with h_2 small enough (but independent of the particular solution u^H) hence, (3.55) is established for all $h \leq \min\{h_1, h_2\}$. The estimate (3.56) is deduced by substituting (3.55) into (3.59) with $\mu = 1$. We postpone the proof of the uniqueness of u^H to the end of Section 3.4.1.3.

FEM with numerical quadrature for indefinite linear problems. The Step 3 in the proof of Theorem 3.4.1 relies on a priori estimates for FEM with numerical quadrature for indefinite linear elliptic problems of the form

$$-\nabla \cdot (\alpha(x)\nabla\varphi(x)) + \beta(x) \cdot \nabla\varphi(x) + \gamma(x)\varphi(x) = f(x) \text{ on } \Omega, \quad \varphi = 0 \text{ on } \partial\Omega. \tag{3.61}$$

This result, which may be of independent interest, generalizes to the case of numerical quadrature a result of Schatz [Sch74]. The proof of the Proposition 3.4.2 below, relies on the Aubin-Nitche duality argument (applied with the adjoint of (3.61)), the Fredholm alternative, and the compact injection of $L^2(\Omega)$ into $H^{-1}(\Omega)$. Note that the bilinear form associated to (3.61) is not uniformly coercive (it is indefinite) but it satisfies the Gårding inequality (for bounded data α, β, γ).

Proposition 3.4.2 *Let $\ell \geq \ell' \geq 1$. Consider the linear problem (3.61) where $\alpha \in (W^{\ell',\infty}(\Omega))^{d\times d}$, $\beta \in (W^{\ell',\infty}(\Omega))^d$, $\gamma \in W^{\ell',\infty}(\Omega)$. Assume (**Q1**), (**Q2**). Assume that the tensor α is uniformly coercive and bounded, i.e. satisfies (3.52). Assume that for all right-hand side in $H^{-1}(\Omega)$, the solution $\varphi \in H^1_0(\Omega)$ of problem (3.61) is unique. For a fixed f, assume that the solution of (3.61) exists with regularity $\varphi \in H^{\ell'+1}(\Omega)$. Then, for all h small enough, the FE problem: find $\varphi^H \in S^\ell_0(\Omega, \mathcal{T}_H)$ such that*

$$(\alpha\nabla\varphi^H, \nabla v^H)_H + (\beta \cdot \nabla\varphi^H + \gamma\varphi^H, v^H)_H = (f, v^H), \quad \forall v^H \in S^\ell_0(\Omega, \mathcal{T}_H), \tag{3.62}$$

possesses a unique solution $\varphi^H \in S^\ell_0(\Omega, \mathcal{T}_H)$; and φ^H satisfies the estimate

$$\|\varphi - \varphi^H\|_{H^1(\Omega)} \leq CH^{\ell'}\|\varphi\|_{H^{\ell'+1}(\Omega)} \tag{3.63}$$

where C is independent of H. [9]

Using the Aubin-Nitsche duality argument for coercive linear elliptic problems with numerical quadrature [CR72, Thm. 10], and assuming additional regularity on the coefficients (e.g. $\alpha_{mn}, \beta_m, \gamma \in W^{\ell'+1,\infty}(\Omega)$), it is also possible to show the optimal L^2 estimate $\|\varphi - \varphi^H\|_{L^2(\Omega)} \leq Ch^{\ell'+1}\|\varphi\|_{H^{\ell'+1}(\Omega)}$.

3.4.1.3 The Newton method and the uniqueness of the numerical solution

We show that under the hypotheses of Theorem 3.4.1, the Newton method (3.64) can be used to compute the numerical solution u^H of the nonlinear system (3.54). Given an initial guess $z_0^H \in S_0^\ell(\Omega, \mathcal{T}_H)$, the Newton method reads

$$N_H(z_k^H; z_{k+1}^H - z_k^H, v^H) = F_H(v^H) - (a(z_k^H)\nabla z_k^H, \nabla v^H)_H, \quad \forall v^H \in S_0^\ell(\Omega, \mathcal{T}_H), \tag{3.64}$$

where

$$N_H(z^H; v^H, w^H) := (a(\cdot, z^H)\nabla v^H, \nabla w^H)_H + (v^H \partial_u a(\cdot, z^H)\nabla z^H, \nabla w^H)_H.$$

Consider for all h the quantity $\sigma_H := \sup_{v^H \in S_0^\ell(\Omega,\mathcal{T}_H)} \|v^H\|_{L^\infty(\Omega)}/\|v^H\|_{H^1(\Omega)}$. Using (3.53), on can show the estimates $\sigma_H \leq C(1+|\ln h|)^{1/2}$ for $d = 2$, and $\sigma_H \leq Ch^{-1/2}$ for $d = 3$, where C is independent of h. Theorem 3.4.3 generalizes the results in [DD75] to the case of numerical quadrature. Its proof uses similar arguments.

Theorem 3.4.3 *Consider u^H a solution of* (3.54). *Under assumptions of Theorem 3.4.1, there exist $H_0, \delta > 0$ such that if $H \leq H_0$ and $\sigma_H\|z_0^H - u^H\|_{H^1(\Omega)} \leq \delta$, then the sequence (z_k^H) for the Newton method* (3.64) *is well defined, and $e_k := \|z_k^H - u^H\|_{H^1(\Omega)}$ is a decreasing sequence that converges quadratically to* 0 *for $k \to \infty$,*

$$e_{k+1} \leq C\sigma_H e_k^2, \tag{3.65}$$

where C is a constant independent of H, k.

Proof of the uniqueness of the FE solution (Theorem 3.4.1). Given two solutions $u^H, \widetilde{u}^H$ of (3.54), we consider the Newton method with initial value $z_0^H = \widetilde{u}^H$. Then, on one hand, $z_k^H = z_0^H$ for all k (as $\widetilde{u}^H$ solves (3.54)). On the other hand, $\sigma_H\|\widetilde{u}^H - u^H\|_{H^1(\Omega)} \leq C\sigma_H h^\ell \to 0$ (as both $\widetilde{u}^H, u^H$ satisfy (3.55)). Hence, Theorem 3.4.3 shows $u^H = \widetilde{u}^H$ for all $H \leq H_3$, where H_3 is small enough. [10] □

3.4.2 The multiscale case: analysis of the nonlinear FE-HMM

We consider the class of nonlinear non-monotone multiscale problems (3.2) together with the corresponding homogenized problem (3.7), as shown in [BM81, Theorem 3.6].

The FE-HMM method for computing a numerical approximation u^H of u_0, essentially similar to the method proposed in [EMZ05][11] reads as follows. It is based on a macroscopic

[9] A quadrature formula can also be used in the right-hand side of (3.62), if f has the regularity of Thm. 3.4.1.

[10] Observe that $H_3 \leq \min\{H_1, H_2\}$, where H_1, H_2 are defined in the proof of the H^1-L^2 bounds, thus $H_0 = H_3$ in Thm. 3.4.1.

[11] In [EMZ05] (3.66) is based on exact micro functions v_{K_j}, w_{K_j} instead of the FE micro functions $v_{K_j}^{h,s}, w_{K_j}^{h,s}$ and the micro-problems are nonlinear (see [EMZ05, equs. (5.3)-(5.4)]).

FEM defined on QF with a macro FE space $S_0^\ell(\Omega, \mathcal{T}_H)$ (defined as in Sect. 3.2), and microscopic FEMs recovering the missing macroscopic tensor at the macroscopic quadrature points. For each macro element $K \in \mathcal{T}_H$ and each integration point $x_{K_j} \in K$, $j = 1, \dots, J$, we define the sampling domains $K_{\delta_j} = x_{K_j} + (-\delta, \delta)^d$, $(\delta \geq \varepsilon)$. For each K_{δ_j}, we then define a micro FE space $S^q(K_{\delta_j}, \mathcal{T}_h) \subset W(K_{\delta_j})$ with simplicial or quadrilateral FEs and a conformal and shape regular family of triangulation $\mathcal{T}_h$. The space $W(K_{\delta_j})$ is either the Sobolev space $W(K_{\delta_j}) = W^1_{per}(K_{\delta_j}) = \{z \in H^1_{per}(K_{\delta_j});\ \int_{K_{\delta_j}} z dx = 0\}$ for a periodic coupling or $W(K_{\delta_j}) = H^1_0(K_{\delta_j})$ for a coupling through Dirichlet boundary conditions.

FE-HMM We define the nonlinear FE-HMM as follows. Find $u^H \in S_0^\ell(\Omega, \mathcal{T}_H)$ such that

$$B_H(u^H; u^H, w^H) = F_H(w^H), \forall w^H \in S_0^\ell(\Omega, \mathcal{T}_H),$$

where

$$B_H(u^H; v^H, w^H) := \sum_{K \in \mathcal{T}_H} \sum_{j=1}^{J} \frac{\omega_{K_j}}{|K_{\delta_j}|} \int_{K_{\delta_j}} a^\varepsilon(x, u^H(x_{K_j})) \nabla v_{K_j}^{h,u^H(x_{K_j})}(x) \cdot \nabla w_{K_j}^{h,u^H(x_{K_j})}(x) dx, \tag{3.66}$$

and $w_{K_j}^{h,u^H(x_{K_j})}$ (and similarly for $v_{K_j}^{h,u^H(x_{K_j})}$) denotes the solution of the micro problem (3.15) with scalar parameter $t = u^H(x_{K_j})$.

3.4.2.1 Fully-discrete a priori error analysis

Consider the following smoothness and structure assumptions (**H2**), (**H3**) on the tensor, analogous to those described in Section 3.3.2 and motivated in Remark 3.3.5.

(**H2**) Given $q \in \mathbb{N}$, the cell functions $\psi_{K_j}^{i,s} \in W(K_{\delta_j})$ of problem (3.21) satisfy the bound $|\psi_{K_j}^{i,s}|_{H^{q+1}(K_{\delta_j})} \leq C\varepsilon^{-q}\sqrt{|K_{\delta_j}|}$, with C independent of ε, the quadrature point x_{K_j}, the domain K_{δ_j}, and the parameter s for all $i = 1 \dots d$. Here, $\mathbf{e}_1, \dots, \mathbf{e}_d$ is the canonical basis of $\mathbb{R}^d$. The same assumption is also made with the tensor a^ε replaced by $(a^\varepsilon)^T$ in (3.21).

(**H3**) For all $m, n = 1, \dots, d$, we assume $a^\varepsilon_{mn}(x, s) = a_{mn}(x, x/\varepsilon, s)$, where $a_{mn}(x, y, s)$ is y-periodic in Y, and the map $(x, s) \mapsto a_{mn}(x, \cdot, s)$ is Lipschitz continuous and bounded from $\overline{\Omega} \times \mathbb{R}$ into $W^{1,\infty}_{per}(Y)$.

Following the framework of analysis presented in Sect. 3.4.1.2 in the context of one-scale problems, we obtain the following H^1 and L^2 a priori estimates of the nonlinear FE-HMM involving the macro and micro mesh sizes H, h.

Theorem 3.4.4 *Let $\ell \geq 1$, $q \geq 1$ and $\mu = 0$ or 1. In addition to the assumptions of Theorem 3.4.1 on problem* (3.7)*, assume (***H2***), (***H3***), and assume that a^ε satisfies* (3.51), (3.52)*. Then, there exist $H_0 > 0$ and $r_0 > 0$ such at if $H \leq H_0$ and $h/\varepsilon \leq r_0$ then*

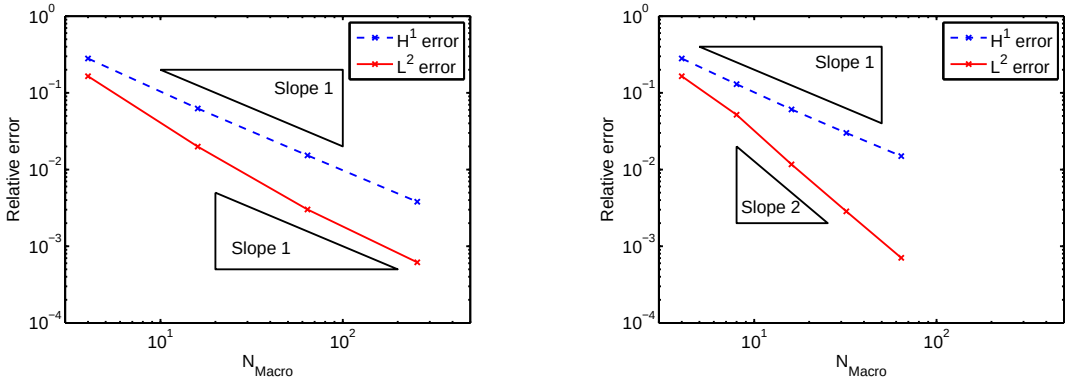
$$\|u_0 - u^H\|_{H^{1-\mu}(\Omega)} \leq \begin{cases} C(H^{\ell+\mu} + (\frac{h}{\varepsilon})^{2q} + \delta), & \text{if } W(K_{\delta_j}) = W^1_{per}(K_{\delta_j}), \delta/\varepsilon \in \mathbb{N}^*, \\ C(H^{\ell+\mu} + (\frac{h}{\varepsilon})^{2q}), & \text{if } W(K_{\delta_j}) = W^1_{per}(K_{\delta_j}),\ \delta/\varepsilon \in \mathbb{N}^*, \text{ and } a^\varepsilon(x,s) \text{ is replaced by } a(x_{K_j}, x/\varepsilon, s) \text{ in (3.66), (3.15), (3.21),} \\ C(H^{\ell+\mu} + (\frac{h}{\varepsilon})^{2q} + \delta + \frac{\varepsilon}{\delta}), & \text{if } W(K_{\delta_j}) = H^1_0(K_{\delta_j})\ (\delta > \varepsilon), \end{cases}$$

where we also assume $\delta \leq r_0$ or $\delta + \varepsilon/\delta \leq r_0$ in the first and third cases, respectively. We use the notation $H^0(\Omega) = L^2(\Omega)$. The constants C are independent of $H, h, \varepsilon, \delta$.

If in addition to the assumptions of Theorem 3.4.4, the map $s \in \mathbb{R} \mapsto a^\varepsilon(\cdot, s) \in (W^{1,\infty}(\Omega))^d$ is of class C^2 with first and second derivatives bounded by $C\varepsilon^{-1}$, then for sufficiently fine meshes and modeling errors (e.g. in the second case of Theorem 3.4.4, for $(h/\varepsilon)^{2q} \leq H \leq H_1$), one can show the convergence of the Newton method used in practice to compute the FE-HMM solution u^H, and the uniqueness of this numerical solution.

3.4.2.2 Numerical examples

We shall illustrate the sharpness of the H^1 and L^2 a priori error estimates of Theorem 3.4.4. First, we consider a simple test problem where the exact homogenized tensor and the exact solution are known analytically. Second, we apply our multiscale method to a steady state model of Richards equation for porous media flows.



(a) Optimal H^1 refinement strategy with $N_{Micro} \sim \sqrt{N_{Macro}}$ where $N_{Micro} = 4, 8, 16, 32$, $N_{Macro} = 4, 16, 64, 256$ respectively.

(b) Optimal L^2 refinement strategy with $N_{Micro} = N_{Macro} = 4, 8, 16, 32, 64$.

Figure 3.1: Nonlinear homogenization test problem (3.7)-(3.68). e_{L^2} error (solid lines) and e_{H^1} error (dashed lines) as a function of the size N_{Macro} of the uniform mesh with $M_{Macro} = N_{Macro} \times N_{Macro}$ $\mathcal{Q}^1$-quadrilateral elements.

Convergence rates: test problem We recall that for a tensor of the form $a^\varepsilon(x, s) = a(x, x/\varepsilon, s)$ where $a(x, y, s)$ is periodic with respect to the fast variable y and collocated in the slow variable x (i.e. (3.66) is used), the H^1 and L^2 errors satisfy (see the second case in Theorem 3.4.4 with $\ell = q = 1$)

$$\|u^H - u_0\|_{H^1(\Omega)} \leq C(H + \hat{h}^2), \qquad \|u^H - u_0\|_{L^2(\Omega)} \leq C(H^2 + \hat{h}^2), \tag{3.67}$$

where $\hat{h} := h/\varepsilon$ is the scaled micro mesh size. In the above estimates, periodic boundary conditions are used for (3.15) and we assume that the micro sampling domains cover one period of the oscillating tensor in each spatial dimension. For rectangular elements, we consider the Gauss quadrature with $J = 4$ nodes $(1/2 \pm \sqrt{3}/6, 1/2 \pm \sqrt{3}/6)$, while for triangular elements, we consider the quadrature formula with $J = 1$ node located at the barycenter. Note that we obtain similar results when considering either rectangular or triangular elements.

We consider the non-linear problem (3.7) on the domain $\Omega = (0, 1)^2$ with homogeneous

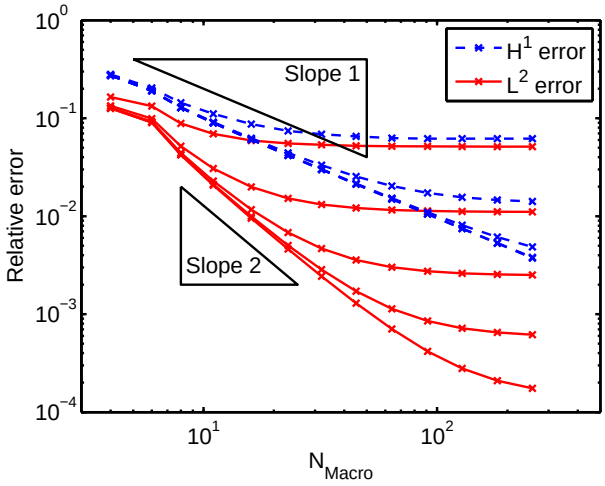


Figure 3.2: Nonlinear homogenization test problem (3.7)-(3.68). e_{L^2} error (solid lines) and e_{H^1} error (dashed lines) as a function of the size N_{Macro} of the uniform mesh with $M_{Macro} = N_{Macro} \times N_{Macro}$ $\mathcal{Q}^1$-quadrilateral elements. The lines correspond respectively to $N_{Micro} = 4, 8, 16, 32, 64$.

Dirichlet boundary conditions and the following anisotropic oscillatory tensor

$$a^\varepsilon(x,s) = \frac{1}{\sqrt{3}} \begin{pmatrix} (2+\sin(2\pi x_1/\varepsilon))(1+x_1\sin(\pi s)) & 0 \\ 0 & (2+\sin(2\pi x_2/\varepsilon))(2+\arctan(s)) \end{pmatrix}. \tag{3.68}$$

The homogenized tensor can be computed analytically and is given by

$$a^0(x,s) = \begin{pmatrix} 1+x_1\sin(\pi s) & 0 \\ 0 & 2+\arctan(s) \end{pmatrix}.$$

The source $f(x)$ in (3.7) is adjusted analytically so that the homogenized solution u_0 is

$$u_0(x) = 8\sin(\pi x_1)x_2(1-x_2), \tag{3.69}$$

The H^1 and L^2 relative errors between the exact homogenized solution u_0 and the FE-HMM solution u^H can be estimated by quadrature with

$$\begin{aligned} e_{L^2}^2 &:= \|u_0\|_{L^2(\Omega)}^{-2} \sum_{K\in\mathcal{T}_H} \sum_{j=1}^{J} \omega_{K_j} |u^H(x_{K_j}) - u_0(x_{K_j})|^2, \\ e_{H^1}^2 &:= \|\nabla u_0\|_{L^2(\Omega)}^{-2} \sum_{K\in\mathcal{T}_H} \sum_{j=1}^{J} \omega_{K_j} \|\nabla u^H(x_{K_j}) - \nabla u_0(x_{K_j})\|^2, \end{aligned}$$

so that

$$e_{L^2} \approx \frac{\|u_0 - u^H\|_{L^2(\Omega)}}{\|u_0\|_{L^2(\Omega)}}, \qquad e_{H^1} \approx \frac{\|\nabla(u_0 - u^H)\|_{L^2(\Omega)}}{\|\nabla u_0\|_{L^2(\Omega)}}.$$

We emphasize that ε is needed for the algorithm but its precise value is not important, as for locally periodic problem solved with periodic micro boundary conditions, the convergence rate and the computational cost are *independent* of ε (see (3.67)). We consider a sequence of uniform macro partitions $\mathcal{T}_H$ with meshsize $H = 1/N_{Macro}$ and

$N_{Macro} = 4, 6, 8, \ldots, 256$. We choose $\mathcal{Q}^1$-rectangular elements with size $H = 1/N_{Macro}$ in the experiments below; the results are silimar for $\mathcal{P}^1$-triangular elements.

In Figure 3.1 the H^1 an L^2 relative errors between the exact homogenized solution and the FE-HMM solutions are shown for the above sequence of partitions using a simultaneous refinement of H and $\hat{h}$ according to $\hat{h} \sim H$ (L^2 norm) and $\hat{h} \sim \sqrt{H}$ (H^1 norm). We observe the expected (optimal) convergence rates (3.67) in agreement with Theorem 3.4.1.

We next show that the ratio between the macro and micro meshes is sharp. For that, we refine the macromesh H while keeping fixed the micro mesh size. This is illustrated in Figure 3.2, where we plot the H^1 an L^2 relative errors as a function of $H = 1/N_{Macro}$. Five sizes of micro meshes are chosen with size $\hat{h}, = 1/N_{Micro}$ and $N = Micro = 4, 8, 16, 32$. We observe that for small values of $H = 1/N_{Macro}$, the error due to the macro domain discretization is dominant. [12] For large values of $N_{Macro} = 1/H$, the error due to the micro domains discretization becomes dominant and the H^1 and L^2 errors becomes independent of N_{Macro} (horizontal lines). We observe that when $N_{Micro} = 1/\hat{h}$ is multiplied by 2, both the H^1 and L^2 errors are divided by 4, which corroborates Theorem 3.4.4 : the micro error has size $\mathcal{O}(\hat{h}^2)$. This experiments illustrate that *simultaneous* refinement of macro and micro meshes (at the right ratio) is needed for optimal convergence rates with minimal computational cost.

Richards equation for multiscale porous media We consider the Richards equation for describing the fluid pressure $u(x,t)$ in an unsaturated porous medium, with multiscale permeability tensor K^ε and volumetric water content Θ^ε,

$$\frac{\partial \Theta^\varepsilon(u_\varepsilon(x))}{\partial t} - \nabla \cdot (K^\varepsilon(u_\varepsilon(x))\nabla u_\varepsilon(x))) + \frac{\partial K^\varepsilon(u_\varepsilon(x))}{\partial x_2} = f(x) \text{ in } \Omega,$$

where x_2 is the vertical coordinate, and f corresponds to possible sources or sinks. We choose an exponential model for the permeability tensor K^ε similar to the one in [CDY05, Sect. 5.1],

$$K^\varepsilon(x,s) = \alpha^\varepsilon(x) e^{\alpha^\varepsilon(x)s} \quad \text{where } \alpha^\varepsilon(x) = \frac{1/117.4}{(2 + 1.8\sin(2\pi(2x_2/\varepsilon - x_1/\varepsilon)))}. \tag{3.70}$$

For our numerical simulation, we consider the steady state $\partial\Theta^\varepsilon(u_\varepsilon)/\partial t = 0$.

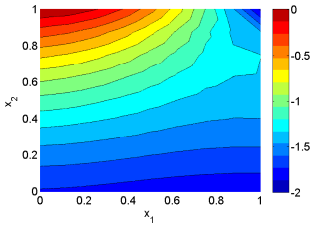
$$-\nabla \cdot (K^\varepsilon(u_\varepsilon(x))\nabla(u_\varepsilon(x) - x_2)) = 0 \text{ in } \Omega = (0,1)^2, \tag{3.71}$$

where for simplicity we set $f(x) \equiv 0$. Note that (3.71) can be cast in the form (3.7) by considering the change of variable $v_\varepsilon(x) = u_\varepsilon(x) - x_2$. We add mixed boundary conditions of Dirichlet and Neumann types. We put Neumann conditions on the left, right and bottom boundaries of the domain (n denotes the vector normal to the boundary) and a Dirichlet condition on the top boundary. Precisely, we take

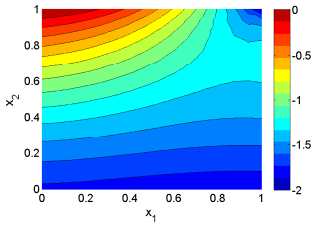
$$\begin{aligned} &u_\varepsilon(x) = -1.9x_1^2 \text{ on } \partial\Omega_D = [0,1] \times \{1\}, \\ &n \cdot (K^\varepsilon(u_\varepsilon(x))\nabla(u_\varepsilon(x) - x_2)) = 0 \text{ on } \partial\Omega_N = \{0,1\} \times [0,1] \cup [0,1] \times \{0\}. \end{aligned}$$

We refine the macro and micro meshes according to the optimal strategy as seen in the above test problem. The numerical results are compared to a resolved standard FE solution for the fine scale problem where $\varepsilon = 10^{-2}$ using $\sim 10^6$ degrees of freedom, and plotted in Fig. 3.3(f). As we compare the fine scale solution with the FE-HMM solution

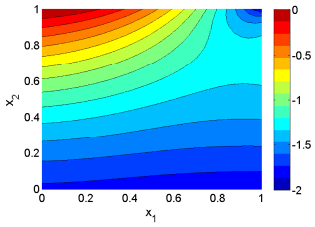
[12] Note that the curves for the H^1 error are nearly identical for $N_{Micro} = 32, 64$.



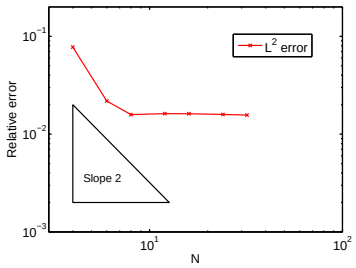
(a) FE-HMM. macro and micro meshes of size 8×8.



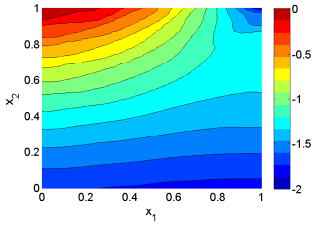
(b) FE-HMM. macro and micro meshes of size 16×16.



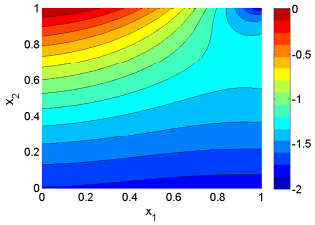
(c) FE-HMM. macro and micro meshes of size 32×32.



(d) FE-HMM. L^2 relative error.



(e) FEM. mesh size: 32×32 (unresolved).



(f) FEM. mesh size: 1024×1024 (finescale).

Figure 3.3: Richards problem (3.70)-(3.71). Top pictures: level curves of the FE-HMM solutions with $N_{Macro} = N_{Micro}$. Fig. (d): L^2 relative error for $u^H - u_\varepsilon$ versus $N = N_{Macro} = N_{Micro}$ (optimal L^2 refinement strategy). Figs. (e)-(f): level curves of the standard FEM solutions.

(without reconstruction) we restrict ourselves to comparison in the L^2 norm. From the results in Sections 3 and 4 we know that

$$\|u^H - u_\varepsilon\|_{L^2(\Omega)} \leq C(H^2 + \hat{h}^2) + \eta_\varepsilon$$

where $\eta_\varepsilon := \|u_0 - u_\varepsilon\|_{L^2(\Omega)} \to 0$ for $\varepsilon \to 0$. We first see in Figure 3.3(d) the expected convergence rate for the L^2 error when macro and micro meshes are refined at the same speed $N_{Macro} = N_{Micro} = N$, and the horizontal line corresponds to the term η_ε, which numerically appears to be of the size[13] of ε. In Figures 3.3(a)-(c), we plot the level curves of the FE-HMM solution for problem (3.50), where we consider uniform $N \times N$ macro meshes with couples of $\mathcal{P}^1$-triangular FEs, and uniform $N \times N$ micro meshes with $\mathcal{Q}^1$-rectangular FEs. For comparison, we also plot the standard FEM solution of (3.7) with a coarse 32×32 mesh (unresolved) and a finescale solution on a fine 1024×1024 mesh. We observe that the unresolved FEM does not yield a qualitative correct result. In contrast, the FE-HMM permits to capture the correct behavior of the resolved solution at a much lower computational cost.

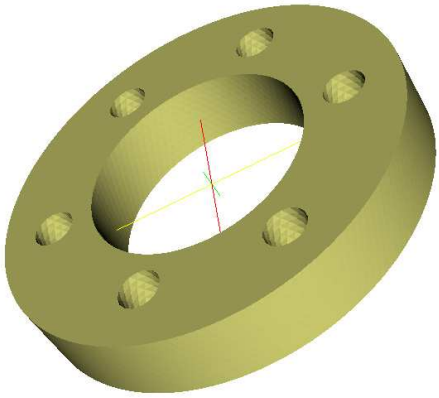
3.5 Perspectives

The framework of the FE-HMM permits in principle to couple the macro FE method with various types of micro FE methods. In [ABV13b, ABV13a] we investigate the coupling with the reduced basis finite element method for the class of quasilinear problems (3.2). This yields the reduced basis finite element heterogeneous multiscale method (RB-FE-HMM). The analysis supported by numerical experiments in 2D and 3D show how the use

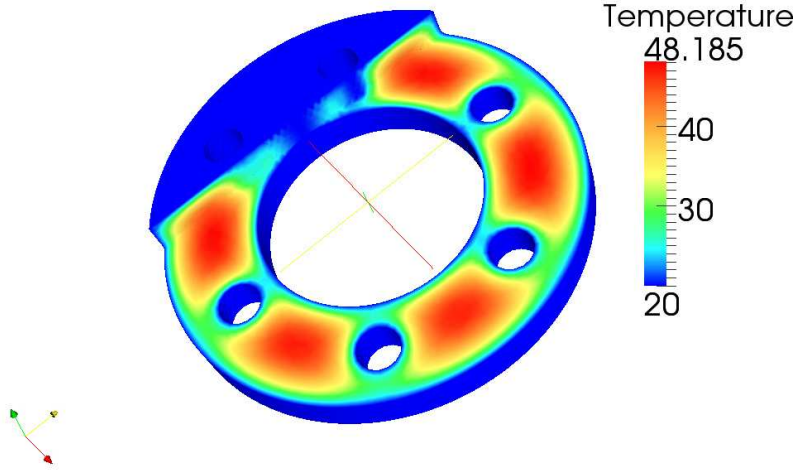
[13] Recall that for linear homogenization problems, one has $\|u_0 - u_\varepsilon\|_{L^2(\Omega)} \leq C\varepsilon$ [JKO94, Sect. 1.4].

of reduced basis (RB) considerably improves the efficiency by orders of magnitude it terms of computational cost, by reducing drastically the number of degrees of freedom especially in 3D (see the example in Fig. 3.4).

In [AHV13] (in preparation), our aim is to extend and analysis the FE-HMM method to parabolic nonlinear problems where the tensor a^ε depends nonlinearly on the gradient ∇u_ε of the solution. Such models arise for instance in multiscale composite carbon materials in magnetostatics [NSDG11]. Due to the oscillatory nature of u_ε, nonlinearities naturally arise in the micro cell problems of homogenization theory which are particularly challenging to treat, both from the numerical and theoretical points of view.



(a) 3D geometry of the computational domain



(b) Nonlinear P1 RB-FE-HMM solution (transversal cut) on a macro mesh with 90000 tetrahedra.

Figure 3.4: Example of the heat transfert in a 3D car rotor part made with composite material. Quasilinear multiscale model of the form (3.2). Figure from [ABV13a].

Bibliography

[AB05] G. Allaire and R. Brizzi. A multiscale finite element method for numerical homogenization. *SIAM, Multiscale Model. Simul.*, 4(3):790–812 (electronic), 2005.

[Abd02] A. Abdulle. Fourth order Chebyshev methods with recurrence relation. *SIAM J. Sci. Comput.*, 23(6):2041–2054, 2002.

[Abd05] A. Abdulle. On a priori error analysis of fully discrete heterogeneous multiscale FEM. *SIAM, Multiscale Model. Simul.*, 4(2):447–459, 2005.

[Abd09] A. Abdulle. The finite element heterogeneous multiscale method: a computational strategy for multiscale pdes. *GAKUTO Int. Ser. Math. Sci. Appl.*, 31:135–184, 2009.

[Abd11] A. Abdulle. A priori and a posteriori error analysis for numerical homogenization: a unified framework. *Ser. Contemp. Appl. Math. CAM*, 16:280–305, 2011.

[Abd12] A. Abdulle. Discontinuous galerkin finite element heterogeneous multiscale method for elliptic problems with multiple scales. *Math. Comp.*, 81(278):687–713, 2012.

[Abd13] A. Abdulle. Numerical homogenization methods. *to appear in Springer, Encyclopedia of Applied and Computational Mathematics*, 2013.

[ABDe09] A. Abdulle, J. Banasiak, A. Damlamian, and M. S. (eds). *Multiple scales problems in Biomathematics, Mechanics and Physics*, volume 31 of *Gakuto Internat. Ser., Math. Sci. Appl.* Gakkotosho Co. Ltd., Tokyo, 2009.

[ABV13a] A. Abdulle, Y. Bai, and G. Vilmart. An offline-online homogenization strategy to solve quasilinear two-scale problems at the cost of one-scale problems. *Submitted for publication*, 2013.

[ABV13b] A. Abdulle, Y. Bai, and G. Vilmart. Reduced basis finite element heterogeneous multiscale method for quasilinear elliptic homogenization problems. *Submitted for publication*, 2013.

[AC08] A. Abdulle and S. Cirilli. S-ROCK: Chebyshev methods for stiff stochastic differential equations. *SIAM J. Sci. Comput.*, 30(2):997–1014, 2008.

[ACVZ12] A. Abdulle, D. Cohen, G. Vilmart, and K. C. Zygalakis. High weak order methods for stochastic differential equations based on modified equations. *SIAM J. Sci. Comput.*, 34(3):A1800–A1823, 2012.

[AD82] M. Artola and G. Duvaut. Un résultat d'homogénéisation pour une classe de problèmes de diffusion non linéaires stationnaires. *Ann. Fac. Sci. Toulouse Math.*, 4(5):1–28, 1982.

[AE03] A. Abdulle and W. E. Finite difference heterogeneous multi-scale method for homogenization problems. *J. Comput. Phys.*, 191(1):18–39, 2003.

[AEEVE12] A. Abdulle, W. E, B. Engquist, and E. Vanden-Eijnden. The heterogeneous multiscale method. *to appear in Acta Numerica*, 2012.

[AHV13] A. Abdulle, M. Huber, and G. Vilmart. Fully-discrete space-time analysis for parabolic nonlinear monotone single scale and multiscale problems. *In preparation*, 2013.

[AL08] A. Abdulle and T. Li. S-ROCK methods for stiff Ito SDEs. *Commun. Math. Sci.*, 6(4):845–868, 2008.

[AM01] A. Abdulle and A. Medovikov. Second order Chebyshev methods based on orthogonal polynomials. *Numer. Math.*, 90(1):1–18, 2001.

[Arn74] L. Arnold. *Stochastic differential equations: theory and applications*. John Wiley and Sons, New york, 1974.

[AV11] A. Abdulle and G. Vilmart. The effect of numerical integration in the finite element method for nonmonotone nonlinear elliptic problems with application to numerical homogenization methods. *C. R. Math. Acad. Sci. Paris*, 349(19-20):1041–1046, 2011.

[AV12a] A. Abdulle and G. Vilmart. Coupling heterogeneous multiscale FEM with Runge-Kutta methods for parabolic homogenization problems: a fully discrete spacetime analysis. *Math. Models Methods Appl. Sci.*, 22(6):1250002, 40, 2012.

[AV12b] A. Abdulle and G. Vilmart. A priori error estimates for finite element methods with numerical quadrature for nonmonotone nonlinear elliptic problems. *Numer. Math.*, 121(3):397–431, 2012.

[AV13a] A. Abdulle and G. Vilmart. Analysis of the finite element heterogeneous multiscale method for quasilinear elliptic homogenization problems. *to appear in Mathematics of Computations*, 2013.

[AV13b] A. Abdulle and G. Vilmart. PIROCK: a swiss-knife partitioned implicit-explicit orthogonal Runge-Kutta Chebyshev integrator for stiff diffusion-advection-reaction problems with or without noise. *J. Comput. Phys.*, 242:869–888, 2013.

[AVZ12] A. Abdulle, G. Vilmart, and K. Zygalakis. Weak second order explicit stabilized methods for stiff stochastic differential equations. *to appear in SIAM J. Sci. Comput.*, 2012.

[AVZ13a] A. Abdulle, G. Vilmart, and K. Zygalakis. Long-run accuracy of numerical integrators for ergodic SDEs. *In preparation*, 2013.

[AVZ13b] A. Abdulle, G. Vilmart, and K. C. Zygalakis. Mean-square A-stable diagonally drift-implicit integrators with high order for stiff Ito systems of stochastic differential equations with noncommutative noise. *to appear in BIT*, 2013.

[BB91] J. Bear and Y. Bachmat. *Introduction to modelling of transport phenomena in porous media*. Kluwer Academic, Dordrecht, The Netherlands, 1991.

[BB10] X. Blanc and C. L. Bris. Improving on computation of homogenized coefficients in the periodic and quasi-periodic setting. *Netw. Heterog. Media*, 5(1):1–19, 2010.

[BBT04] K. Burrage, P. Burrage, and T. Tian. Numerical methods for strong solutions of stochastic differential equations: an overview. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 460(2041):373–402, 2004.

[BCR99] S. Blanes, F. Casas, and J. Ros. Symplectic integrators with processing: a general study. *SIAM J. Sci. Comput.*, 21:149–161, 1999.

[BJS64] L. Bers, F. John, and M. Schechter. *Partial differential equations*. Lectures in Applied Mathematics, Vol. III. Interscience Publishers John Wiley & Sons, Inc. New York-London-Sydney, 1964.

[BK10] E. Buckwar and C. Kelly. Towards a systematic linear stability analysis of numerical methods for systems of stochastic differential equations. *SIAM Journal on Numerical Analysis*, 48(1):298–321, 2010.

[BLP78] A. Bensoussan, J.-L. Lions, and G. Papanicolaou. *Asymptotic analysis for periodic structures*. North-Holland Publishing Co., Amsterdam, 1978.

[BM81] L. Boccardo and F. Murat. Homogénéisation de problèmes quasi-linéaires. *Publ. IRMA, Lille.*, 3(7):13–51, 1981.

[BOFM92] S. Brahim-Otsmane, G. A. Francfort, and F. Murat. Correctors for the homogenization of the wave and heat equations. *J. Math. Pures Appl. (9)*, 71(3):197–231, 1992.

[BP04] A. Bourgeat and A. Piatnitski. Approximations of effective coefficients in stochastic homogenization. *Ann. Inst. H. Poincaré Probab. Statist.*, 40(2):153–165, 2004.

[Bro00] C. Brouder. Runge-Kutta methods and renormalization. *Euro. Phys. J. C*, 12:521–534, 2000.

[Bro04] C. Brouder. Trees, renormalization and differential equations. *BIT*, 44(3):425–438, 2004.

[BS08] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.

[But69] J. C. Butcher. The effective order of Runge-Kutta methods. In J. L. Morris, editor, *Proceedings of Conference on the Numerical Solution of Differential Equations*, volume 109 of *Lecture Notes in Math.*, pages 133–139, 1969.

[But72] J. C. Butcher. An algebraic theory of integration methods. *Math. Comput.*, 26:79–106, 1972.

[But08] J. C. Butcher. *Numerical methods for ordinary differential equations*. John Wiley & Sons Ltd., Chichester, second edition, 2008.

[Cay57] A. Cayley. On the theory of the analytic forms called trees. *Phil. Magazine XIII*, pages 172–176, 1857.

[CCDV09] F. Castella, P. Chartier, S. Descombes, and G. Vilmart. Splitting methods with complex times for parabolic equations. *BIT*, 49(3):487–508, 2009.

[CCMSS11] M. P. Calvo, P. Chartier, A. Murua, and J. M. Sanz-Serna. Numerical stroboscopic averaging for ODEs and DAEs. *Appl. Numer. Math.*, 61(10):1077–1095, 2011.

[CD99] D. Cioranescu and P. Donato. *An introduction to homogenization*, volume 17 of *Oxford Lecture Series in Mathematics and its Applications*. The Clarendon Press Oxford University Press, New York, 1999.

[CDY05] Z. Chen, W. Deng, and H. Ye. Upscaling of a class of nonlinear parabolic equations for the flow transport in heterogeneous porous media. *Commun. Math. Sci.*, 3(4):493–515, 2005.

[CEFM09] D. Calaque, K. Ebrahimi-Fard, and D. Manchon. Two interacting Hopf algebras of trees. *to appear in Adv. in Appl. Math.*, 2009.

[Chi09] M. Chipot. *Elliptic equations: an introductory course*. Birkhäuser Advanced Texts: Basler Lehrbücher. [Birkhäuser Advanced Texts: Basel Textbooks]. Birkhäuser Verlag, Basel, 2009.

[CHV05] P. Chartier, E. Hairer, and G. Vilmart. A substitution law for B-series vector fields. *INRIA Report, No. 5498*, 2005.

[CHV07a] P. Chartier, E. Hairer, and G. Vilmart. Modified differential equations. In *Journées d'Analyse Fonctionnelle et Numérique en l'honneur de Michel Crouzeix*, volume 21 of *ESAIM Proc.*, pages 16–20. EDP Sci., Les Ulis, 2007.

[CHV07b] P. Chartier, E. Hairer, and G. Vilmart. Numerical integrators based on modified differential equations. *Math. Comp.*, 76(260):1941–1953 (electronic), 2007.

[CHV09] M. Chyba, E. Hairer, and G. Vilmart. The role of symplectic integrators in optimal control. *Optimal Control Appl. Methods*, 30(4):367–382, 2009.

[CHV10] P. Chartier, E. Hairer, and G. Vilmart. Algebraic structures of B-series. *Found. Comput. Math.*, 10(4):407–427, 2010.

[Cia91] P. G. Ciarlet. Basic error estimates for elliptic problems. In *Handbook of numerical analysis, Vol. II*, Handb. Numer. Anal., II, pages 17–351. North-Holland, Amsterdam, 1991.

[CJMR04] M. Calvo, L. O. Jay, J. I. Montijano, and L. Rández. Approximate compositions of a near identity map by multi-revolution Runge-Kutta methods. *Numer. Math.*, 97(4):635–666, 2004.

[CK98] A. Connes and D. Kreimer. Hopf algebras, renormalization and noncommutative geometry. *Commun. Math. Phys.*, 199(1):203–242, 1998.

[CK00] A. Connes and D. Kreimer. Renormalization in quantum field theory and the Riemann–Hilbert problem. I. the Hopf algebra structure of graphs and the main theorem. *Commun. Math. Phys.*, 210(1):249–273, 2000.

[CM98] A. Connes and H. Moscovici. Hopf algebras, cyclic cohomology and the transverse index theorem. *Comm. Math. Phys.*, 198(1):199–246, 1998.

[CMMV13] P. Chartier, J. Makazaga, A. Murua, and G. Vilmart. Multi-revolution composition methods for highly oscillatory problems. *Submitted for publication*, 2013.

[CMR03] M. Calvo, J. I. Montijano, and L. Rández. A family of explicit multirevolution Runge-Kutta methods of order five. In *Analytic and numerical techniques in orbital dynamics (Spanish) (Albarracín, 2002)*, Monogr. Real Acad. Ci. Exact. Fís.-Quím. Nat. Zaragoza, 22, pages 45–54. Acad. Cienc. Exact. Fís. Quím. Nat. Zaragoza, Zaragoza, 2003.

[CMR07] M. Calvo, J. I. Montijano, and L. Rández. On explicit multi-revolution Runge-Kutta schemes. *Adv. Comput. Math.*, 26(1-3):105–120, 2007.

[CMSS10] P. Chartier, A. Murua, and J. M. Sanz-Serna. Higher-order averaging, formal series and numerical integration I: B-series. *Found. Comput. Math.*, 10(6):695–727, 2010.

[CMSS12] P. Chartier, A. Murua, and J. M. Sanz-Serna. A formal series approach to averaging: exponentially small error estimates. *Discrete Contin. Dyn. Syst.*, 32(9):3009–3027, 2012.

[CR72] P. G. Ciarlet and P.-A. Raviart. The combined effect of curved boundaries and numerical integration in isoparametric finite element methods. In *The mathematical foundations of the finite element method with applications to partial differential equations (Proc. Sympos., Univ. Maryland, Baltimore, Md., 1972)*, pages 409–474. Academic Press, New York, 1972.

[Cro05] M. Crouzeix. Approximation of parabolic equations, 2005. Lecture notes available at `http://perso.univ-rennes1.fr/michel.crouzeix/`.

[CS08] Z. Chen and T. Y. Savchuk. Analysis of the multiscale finite element method for nonlinear and random homogenization problems. *SIAM J. Numer. Anal.*, 46(1):260–279, 2007/08.

[DD75] J. Douglas, Jr. and T. Dupont. A Galerkin method for a nonlinear Dirichlet problem. *Math. Comp.*, 29:689–696, 1975.

[DF12] A. Debussche and E. Faou. Weak backward error analysis for SDEs. *SIAM J. Numer. Anal.*, 50(3):1735–1752, 2012.

[DG00] A. M. Davie and J. G. Gaines. Convergence of numerical schemes for the solution of parabolic stochastic partial differential equations. *Math. Comp.*, 70:121–134, 2000.

[DGS73] E. De Giorgi and S. Spagnolo. Sulla convergenza degli integrali dell'energia per operatori ellittici del secondo ordine. *Boll. Un. Mat. Ital. (4)*, 8:391–411, 1973.

[DR09a] K. Debrabant and A. Rößler. Diagonally drift-implicit runge–kutta methods of weak order one and two for itô sdes and stability analysis. *Appl. Num. Math.*, 59(3–4):595 – 607, 2009.

[DR09b] K. Debrabant and A. Rößler. Families of efficient second order Runge-Kutta methods for the weak approximation of Itô stochastic differential equations. *Appl. Numer. Math.*, 59(3-4):582–594, 2009.

[Dür86] A. Dür. Möbius functions, incidence algebras and power series representations. In *Lecture Notes in Math.*, volume 1202. Springer-Verlag, 1986.

[EE03] W. E and B. Engquist. The heterogeneous multiscale methods. *Commun. Math. Sci.*, 1(1):87–132, 2003.

[EEL+07] W. E, B. Engquist, X. Li, W. Ren, and E. Vanden-Eijnden. Heterogeneous multiscale methods: a review. *Commun. Comput. Phys.*, 2(3):367–450, 2007.

[EMZ05] W. E, P. Ming, and P. Zhang. Analysis of the heterogeneous multiscale method for elliptic homogenization problems. *J. Amer. Math. Soc.*, 18(1):121–156, 2005.

[EP03] Y. Efendiev and A. Pankov. Numerical homogenization of monotone elliptic operators. *Multiscale Model. Simul.*, 2(1):62–79, 2003.

[EP04] Y. Efendiev and A. Pankov. Numerical homogenization and correctors for nonlinear elliptic equations. *SIAM J. Appl. Math.*, 65(1):43–68, 2004.

[Fen86] K. Feng. Difference schemes for Hamiltonian formalism and symplectic geometry. *J. Comp. Math.*, 4:279–289, 1986.

[FJ10] F. Filbet and S. Jin. A class of asymptotic-preserving schemes for kinetic equations and related problems with stiff sources. *J. Comput. Phys.*, 229(20):7625–7648, 2010.

[FKS93] M. Feistauer, M. Křížek, and V. Sobotíková. An analysis of finite element variational crimes for a nonlinear elliptic problem of a nonmonotone type. *East-West J. Numer. Math.*, 1(4):267–285, 1993.

[FŽ87] M. Feistauer and A. Ženíšek. Finite element solution of nonlinear elliptic problems. *Numer. Math.*, 50(4):451–475, 1987.

[Gar88] T. Gard. *Introduction to stochastic differential equations*. Marcel Dekker, New York, 1988.

[Gil08] M. Giles. Multilevel Monte Carlo path simulation. *Operations Research*, 56(3):607–617, 2008.

[Glo06] A. Gloria. An analytical framework for the numerical homogenization of monotone elliptic operators and quasiconvex energies. *Multiscale Model. Simul.*, 5(3):996–1043 (electronic), 2006.

[Glo11] A. Gloria. Reduction of the resonance error—Part 1: Approximation of homogenized coefficients. *Math. Models Methods Appl. Sci.*, 21(8):1601–1630, 2011.

[Glo12] A. Gloria. Numerical approximation of effective coefficients in stochastic homogenization of discrete elliptic equations. *ESAIM Math. Model. Numer. Anal.*, 46(1):1–38, 2012.

[GVB11] B. Grébert and C. Villegas-Blas. On the energy exchange between resonant modes in nonlinear Schrödinger equations. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 28(1):127–134, 2011.

[Has80] R. Hasminskii. *Stochastic stability of differential equations*. Sijthoff and Noordhoff, The Netherlands, 1980.

[HH52] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. of Physiology*, 117(4):500–544, 1952.

[Hig00] D. Higham. A-stability and stochastic mean-square stability. *BIT*, 40:404–409, 2000.

[HLW06] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations*. Springer Series in Computational Mathematics 31. Springer-Verlag, Berlin, second edition, 2006.

[HNW93] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I. Nonstiff Problems*. Springer Series in Computational Mathematics 8. Springer, Berlin, 2 edition, 1993.

[HV06] E. Hairer and G. Vilmart. Preprocessed discrete Moser-Veselov algorithm for the full dynamics of a rigid body. *J. Phys. A*, 39(42):13225–13235, 2006.

[HW74] E. Hairer and G. Wanner. On the Butcher group and general multi-value methods. *Computing*, 13:1–15, 1974.

[HW96] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*. Springer Series in Computational Mathematics 14. Springer-Verlag, Berlin, 2 edition, 1996.

[HWC99] T. Hou, X. Wu, and Z. Cai. Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients. *Math. Comp.*, 68(227):913–943, 1999.

[Jin99] S. Jin. Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations. *SIAM J. Sci. Comput.*, 21(2):441–454, 1999.

[JKO94] V. Jikov, S. Kozlov, and O. Oleinik. *Homogenization of differential operators and integral functionals*. Springer-Verlag, Berlin, Heidelberg, 1994.

[KB13] Y. Komori and K. Burrage. Strong first order S-ROCK methods for stochastic differential equations. *Comput. Appl. Math.*, 242:261–274, 2013.

[Kir88] U. Kirchgraber. An ODE-solver based on the method of averaging. *Numer. Math.*, 53(6):621–652, 1988.

[KL08] A. Karageorghis and D. Lesnic. Steady-state nonlinear heat conduction in composite materials using the method of fundamental solutions. *Comput. Methods Appl. Mech. Engrg.*, 197(33–40):3122–3137, 2008.

[KP92] P. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*. Springer-Verlag, Berlin and New York, 1992.

[KPH95] P. Kloeden, E. Platen, and N. Hofmann. Extrapolation methods for the weak approximation of Itô diffusions. *SIAM J. Numer. Anal.*, 32(5):1519–1534, 1995.

[Lad85] O. A. Ladyzhenskaya. *The boundary value problems of mathematical physics*, volume 49 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1985. Translated from the Russian by Jack Lohwater [Arthur J. Lohwater].

[LCO09] J.-A. Lázaro-Camí and J.-P. Ortega. Reduction, reconstruction, and skew-product decomposition of symmetric stochastic differential equations. *Stoch. Dyn.*, 9(1):1–46, 2009.

[Leb89] V. Lebedev. Explicit difference schemes with time-variable steps for solving stiff systems of equations. *Sov. J. Numer. Anal. Math. Modelling*, 4(2):111–135, 1989.

[Lia97] M. Liao. Random motion of a rigid body. *J. Theoret. Probab.*, 10(1):201–211, 1997.

[LM68] J.-L. Lions and E. Magenes. *Problèmes aux limites non homogènes et applications. Vol. 1*. Travaux et Recherches Mathématiques, No. 17. Dunod, Paris, 1968.

[LM08] M. Lemou and L. Mieussens. A new asymptotic preserving scheme based on micro-macro formulation for linear kinetic equations in the diffusion limit. *SIAM J. Sci. Comput.*, 31(1):334–368, 2008.

[LO95] C. Lubich and A. Ostermann. Runge-Kutta approximation of quasi-linear parabolic equations. *Math. Comp.*, 64(210):601–627, 1995.

[LR04] B. Leimkuhler and S. Reich. *Simulating Hamiltonian Dynamics*. Cambridge Monographs on Applied and Computational Mathematics 14. Cambridge University Press, Cambridge, 2004.

[Mao94] X. Mao. Stochastic stabilization and destabilization. *Systems Control Lett.*, 23(4):279–290, 1994.

[Mar55] G. Maruyama. Continuous markov processes and stochastic equations. *Rend. Circ. Mat. Palermo*, 4:48–90, 1955.

[McL95] R. I. McLachlan. Composition methods in the presence of small parameters. *BIT*, 35:258–268, 1995.

[Mer57] R. H. Merson. An operational method for the study of integration processes. In *Proc. Symp. Data Processing Weapons Research Establishment*, pages 110–1 to 110–25, Salisbury Australia, 1957.

[Mil78] G. N. Milstein. A method of second order accuracy integration of stochastic differential equation. *Theory Probab. Appl.*, 23:396–401, 1978.

[Mil86] G. N. Milstein. Weak approximation of solutions of systems of stochastic differential equations. *Theory Probab. Appl.*, 30(4):750–766, 1986.

[Mil87] G. N. Milstein. A theorem on the order of convergence of mean-square approximations of solutions of systems of stochastic differential equations. *Teor. Veroyatnost. i Primenen.*, 32(4):809–811, 1987.

[MP97] B. Melendo and M. Palacios. A new approach to the construction of multirevolution methods and their implementation. *Appl. Numer. Math.*, 23(2):259–274, 1997.

[MRT02] G. Milstein, Y. Repin, and M. Tretyakov. Numerical methods for stochastic systems preserving symplectic structure. *SIAM J. Numer. Anal.*, 40(4):1583–1604 (electronic), 2002.

[MSS99] A. Murua and J. M. Sanz-Serna. Order conditions for numerical integrators obtained by composing simpler integrators. *Philos. Trans. Royal Soc. London ser. A*, 357:1079–1100, 1999.

[MT97] F. Murat and L. Tartar. H-convergence. In *Topics in the mathematical modelling of composite materials*, volume 31 of *Progr. Nonlinear Differential Equations Appl.*, pages 21–43. Birkhäuser Boston, Boston, MA, 1997.

[MT04] G. N. Milstein and M. V. Tretyakov. *Stochastic numerics for Mathematical Physics*. Scientific Computing. Springer-Verlag, Berlin and New York, 2004.

[MV91] J. Moser and A. P. Veselov. Discrete versions of some classical integrable systems and factorization of matrix polynomials. *Comm. Math. Phys.*, 139:217–243, 1991.

[MZ05] R. I. McLachlan and A. Zanna. The discrete Moser–Veselov algorithm for the free rigid body, revisited. *Found. Comput. Math.*, 5:87–123, 2005.

[MZ07] P. Ming and P. Zhang. Analysis of the heterogeneous multiscale method for parabolic homogenization problems. *Math. Comp.*, 76(257):153–177 (electronic), 2007.

[NSDG11] I. Niyonzima, R. V. Sabariego, P. Dular, and C. Geuzaine. Finite element computational homogenization of nonlinear multiscale materials in magnetostatics. 2011. Proceedings of COMPUMAG-Sydney, 1. Static and quasi-static fields, 8. Material modelling.

[Øks03] B. Øksendal. *Stochastic differential equations*. Universitext. Springer-Verlag, Berlin, sixth edition, 2003. An introduction with applications.

[PJY97] L. R. Petzold, L. O. Jay, and J. Yen. Numerical solution of highly oscillatory ordinary differential equations. In *Acta numerica, 1997*, volume 6 of *Acta Numer.*, pages 437–483. Cambridge Univ. Press, Cambridge, 1997.

[Pla92] E. Platen. High-order weak approximation of ito diffusions by markov chains. *Probab. Engrg. Inform. Sci.*, 6:391–408, 1992.

[PS08] G. A. Pavliotis and A. M. Stuart. *Multiscale methods*, volume 53 of *Texts in Applied Mathematics*. Springer, New York, 2008. Averaging and homogenization.

[RB08] A. Rathinasamy and K. Balachandran. Mean-square stability of second-order Runge-Kutta methods for multi-dimensional linear stochastic differential systems. *J. Comput. Appl. Math.*, 219(1):170–197, 2008.

[Ris89] H. Risken. *The Fokker-Planck equation*, volume 18 of *Springer Series in Synergetics*. Springer-Verlag, Berlin, 1989.

[Röß03] A. Rößler. *Runge-Kutta methods for the numerical solution of stochastic diifferential equations*. Diss. TU Darmstadt. Shaker Verlag, Aachen, 2003.

[Röß09] A. Rößler. Second order Runge-Kutta methods for Itô stochastic differential equations. *SIAM J. Numer. Anal.*, 47(3):1713–1738, 2009.

[Rut83] R. D. Ruth. A canonical integration technique. *IEEE Trans. Nuclear Science*, NS-30:2669–2671, 1983.

[Sch74] A. H. Schatz. An observation concerning Ritz-Galerkin methods with indefinite bilinear forms. *Math. Comp.*, 28:959–962, 1974.

[Sha06] T. Shardlow. Modified equations for stochastic differential equations. *BIT*, 46(1):111–125, 2006.

[SM96] Y. Saito and T. Mitsui. Stability analysis of numerical schemes for stochastic differential equations. *SIAM J. Numer. Anal.*, 33:2254–2267, 1996.

[SM02] Y. Saito and T. Mitsui. Mean-square stability of numerical schemes for stochastic differential systems. *Vietnam J. Math.*, 30(suppl.):551–560, 2002.

[Spa68] S. Spagnolo. Sulla convergenza di soluzioni di equazioni paraboliche ed ellittiche. *Ann. Scuola Norm. Sup. Pisa (3) 22 (1968), 571-597; errata, ibid. (3)*, 22:673, 1968.

[SSV98] B. Sommeijer, L. Shampine, and J. Verwer. RKC: an explicit solver for parabolic PDEs. *J. Comput. Appl. Math.*, 88:316–326, 1998.

[Tal84] D. Talay. Efficient numerical schemes for the approximation of expectations of functionals of the solution of a SDE and applications. *Lecture Notes in Control and Inform. Sci., Springer*, 61:294–313, 1984.

[Toc05] A. Tocino. Mean-square stability of second-order Runge-Kutta methods for stochastic differential equations. *J. Comput. Appl. Math.*, 175(2):355–367, 2005.

[TT90] D. Talay and L. Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Anal. Appl.*, 8(4):483–509 (1991), 1990.

[TVA02] A. Tocino and J. Vigo-Aguiar. Weak second order conditions for stochastic Runge-Kutta methods. *SIAM J. Sci. Comput.*, 24(2):507–523, 2002.

[VHS90] J. Verwer, W. Hundsdorfer, and B. Sommeijer. Convergence properties of the runge-kutta-chebyshev method. *Numer. Math.*, 57:157–178, 1990.

[Vil08a] G. Vilmart. *Étude d'intégrateurs géométriques pour des équations différentielles (in english).* PhD thesis, Univ. Rennes 1/Univ. Genève, 2008. Thesis No. 3758/No. 4038.

[Vil08b] G. Vilmart. Reducing round-off errors in rigid body dynamics. *J. Comput. Phys.*, 227(15):7083–7088, 2008.

[Vil13] G. Vilmart. Rigid body dynamics. *to appear in Springer, Encyclopedia of Applied and Computational Mathematics*, 2013.

[VS80] P. Van der Houwen and B. Sommeijer. On the internal stage Runge-Kutta methods for large m-values. *Z. Angew. Math. Mech.*, 60:479–485, 1980.

[VSH04] J. G. Verwer, B. P. Sommeijer, and W. Hundsdorfer. RKC time-stepping for advection-diffusion-reaction problems. *J. Comput. Phys.*, 201(1):61–79, 2004.

[Wal86] J. Walsh. An introduction to stochastic partial differential equations. In P. L. Hennequin, editor, *École d'Été de Probabilités de Saint Flour XIV - 1984*, volume 1180 of *Lecture Notes in Mathematics*, chapter 3, pages 265–439. Springer Berlin / Heidelberg, 1986.

[WHN09] A. Whittington, A. Hofmeister, and P. Nabelek. Temperature-dependent thermal diffusivity of the earth's crust and implications for magmatism. *Nature*, 458:319–321, 2009.

[Yur86] V. V. Yurinskiĭ. Averaging of symmetric diffusion in a random medium. *Sibirsk. Mat. Zh.*, 27(4):167–180, 215, 1986.

[Zbi11] C. J. Zbinden. Partitioned Runge-Kutta-Chebyshev methods for diffusion-advection-reaction problems. *SIAM J. Sci. Comput.*, 33(4):1707–1725, 2011.

[Zyg11] K. C. Zygalakis. On the existence and the applications of modified equations for stochastic differential equations. *SIAM J. Sci. Comput.*, 33(1):102–130, 2011.

Personal bibliography

[ABV13a] A. Abdulle, Y. Bai, and G. Vilmart. An offline-online homogenization strategy to solve quasilinear two-scale problems at the cost of one-scale problems. *Submitted for publication*, 2013.

[ABV13b] A. Abdulle, Y. Bai, and G. Vilmart. Reduced basis finite element heterogeneous multiscale method for quasilinear elliptic homogenization problems. *Submitted for publication*, 2013.

[ACVZ12] A. Abdulle, D. Cohen, G. Vilmart, and K. C. Zygalakis. High weak order methods for stochastic differential equations based on modified equations. *SIAM J. Sci. Comput.*, 34(3):A1800–A1823, 2012.

[AHV13] A. Abdulle, M. Huber, and G. Vilmart. Fully-discrete space-time analysis for parabolic nonlinear monotone single scale and multiscale problems. *In preparation*, 2013.

[AV11] A. Abdulle and G. Vilmart. The effect of numerical integration in the finite element method for nonmonotone nonlinear elliptic problems with application to numerical homogenization methods. *C. R. Math. Acad. Sci. Paris*, 349(19-20):1041–1046, 2011.

[AV12a] A. Abdulle and G. Vilmart. Coupling heterogeneous multiscale FEM with Runge-Kutta methods for parabolic homogenization problems: a fully discrete spacetime analysis. *Math. Models Methods Appl. Sci.*, 22(6):1250002, 40, 2012.

[AV12b] A. Abdulle and G. Vilmart. A priori error estimates for finite element methods with numerical quadrature for nonmonotone nonlinear elliptic problems. *Numer. Math.*, 121(3):397–431, 2012.

[AV13a] A. Abdulle and G. Vilmart. Analysis of the finite element heterogeneous multiscale method for quasilinear elliptic homogenization problems. *to appear in Mathematics of Computations*, 2013.

[AV13b] A. Abdulle and G. Vilmart. PIROCK: a swiss-knife partitioned implicit-explicit orthogonal Runge-Kutta Chebyshev integrator for stiff diffusion-advection-reaction problems with or without noise. *J. Comput. Phys.*, 242:869–888, 2013.

[AVZ12] A. Abdulle, G. Vilmart, and K. Zygalakis. Weak second order explicit stabilized methods for stiff stochastic differential equations. *to appear in SIAM J. Sci. Comput.*, 2012.

[AVZ13a] A. Abdulle, G. Vilmart, and K. Zygalakis. Long-run accuracy of numerical integrators for ergodic SDEs. *In preparation*, 2013.

[AVZ13b] A. Abdulle, G. Vilmart, and K. C. Zygalakis. Mean-square A-stable diagonally drift-implicit integrators with high order for stiff Ito systems of stochastic differential equations with noncommutative noise. *to appear in BIT*, 2013.

[CCDV09] F. Castella, P. Chartier, S. Descombes, and G. Vilmart. Splitting methods with complex times for parabolic equations. *BIT*, 49(3):487–508, 2009.

[CHV05] P. Chartier, E. Hairer, and G. Vilmart. A substitution law for B-series vector fields. *INRIA Report, No. 5498*, 2005.

[CHV07a] P. Chartier, E. Hairer, and G. Vilmart. Modified differential equations. In *Journées d'Analyse Fonctionnelle et Numérique en l'honneur de Michel Crouzeix*, volume 21 of *ESAIM Proc.*, pages 16–20. EDP Sci., Les Ulis, 2007.

[CHV07b] P. Chartier, E. Hairer, and G. Vilmart. Numerical integrators based on modified differential equations. *Math. Comp.*, 76(260):1941–1953 (electronic), 2007.

[CHV09] M. Chyba, E. Hairer, and G. Vilmart. The role of symplectic integrators in optimal control. *Optimal Control Appl. Methods*, 30(4):367–382, 2009.

[CHV10] P. Chartier, E. Hairer, and G. Vilmart. Algebraic structures of B-series. *Found. Comput. Math.*, 10(4):407–427, 2010.

[CMMV13] P. Chartier, J. Makazaga, A. Murua, and G. Vilmart. Multi-revolution composition methods for highly oscillatory problems. *Submitted for publication*, 2013.

[HV06] E. Hairer and G. Vilmart. Preprocessed discrete Moser-Veselov algorithm for the full dynamics of a rigid body. *J. Phys. A*, 39(42):13225–13235, 2006.

[Vil08a] G. Vilmart. *Étude d'intégrateurs géométriques pour des équations différentielles (in english)*. PhD thesis, Univ. Rennes 1/Univ. Genève, 2008. Thesis No. 3758/No. 4038.

[Vil08b] G. Vilmart. Reducing round-off errors in rigid body dynamics. *J. Comput. Phys.*, 227(15):7083–7088, 2008.

[Vil13] G. Vilmart. Rigid body dynamics. *to appear in Springer, Encyclopedia of Applied and Computational Mathematics*, 2013.

Résumé

Mes travaux de recherche portent sur l'analyse numérique des intégrateurs géométriques et multi-échelles pour les équations différentielles déterministes ou stochastiques. Les modèles d'équations différentielles issus de la physique ou la chimie possèdent souvent une structure géométrique ou multi-échelles particulière (par exemple, les structures hamiltoniennes, les intégrales premières, les structures multi-échelles en temps ou en espace, les systèmes hautement oscillatoires), mais leur complexité est souvent telle qu'une solution satisfaisante est hors de portée en utilisant seulement des méthodes numériques standards à usage général. L'objectif est donc d'identifier les propriétés géométriques ou multi-échelles pertinentes de ces problèmes, et d'en tirer avantage pour concevoir et analyser de nouveaux intégrateurs efficaces, fiables et précis, reproduisant fidèlement le comportement qualitatif de la solution exacte des modèles considérés.

Mots-clés :

équations différentielles ordinaires, équations différentielles stochastiques, équations aux dérivées partielles, intégration numérique géométrique, problèmes raides, systèmes hautement oscillants, méthodes numériques d'homogénéisation, éléments finis.

Summary

My research focuses on the numerical analysis of geometric and multiscale integrators for deterministic or stochastic differential equations. Numerous physical or chemical phenomena can be modeled by differential equations which possess a particular geometric or multiscale structure (e.g. Hamiltonian structures, first integrals, multiscale structures in time or in space, highly oscillatory systems), but their complexity is often so huge that a satisfactory solution is out of reach using only general purpose numerical methods. The aim is thus to identify the relevant geometric or multiscale properties of such problems, and try to take advantage of them to design and study new efficient, reliable, and accurate integrators, that reproduce the qualitative behavior of the exact solution of the considered models.

Keywords:

ordinary differential equations, stochastic differential equations, partial differential equations, geometric numerical integration, stiff problems, highly oscillatory problems, numerical homogenization methods, finite elements.

École normale supérieure de Cachan - Antenne de Bretagne

Campus de Ker Lann - Avenue Robert Schuman - 35170 BRUZ

Tél : +33(0)2 99 05 93 00 - Fax : +33(0)2 99 05 93 29